Data Analytics Bootcamp Project 04

Team 01:
Jaime Bravo
Roberto Rodas..

Context

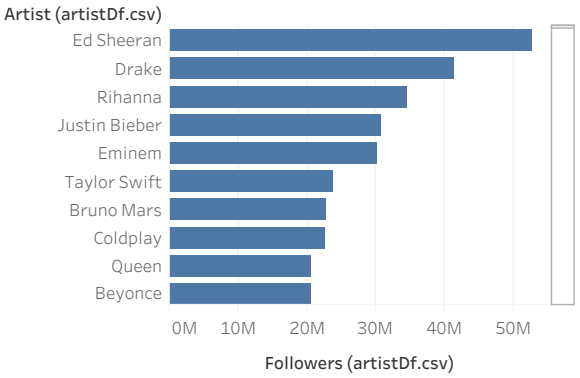
Re-exploring the topic of our first project, we decided to apply more sophisticated tools to develop models that could predict success for a musical project today, considering the vast amounts of data generated by musical streaming and traditional data-collection means, in order to create valuable information for industry executives, musicians, marketing teams and ..

Research Questions

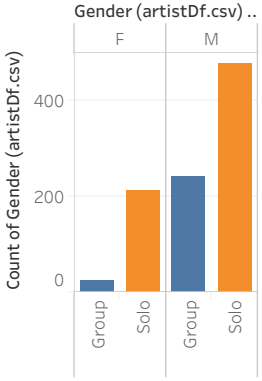
- How are stream numbers related to other success metrics, such as certifications and chart positioning.
- How do song attributes relate to musical success?
- Which labels produce more successful releases?
- Does featuring with another artist raise the success probabilities?

Artist Panorama

We considered it was important to display the general panorama of the data that we worked with, and the element that was at the center of it a..



The data that we worked with comprised all entries into the Billboard Charts from the year 1999 to 2019. In this first table we can see artists ranked by their number o..

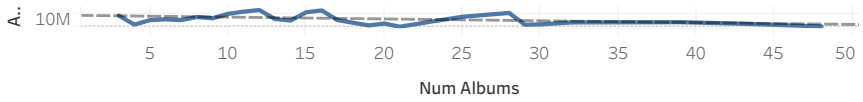


And in this table we can see how the genders, as well as type of aggrupation are distributed. We can see more prevalence of male artists, and a greater tendency ..

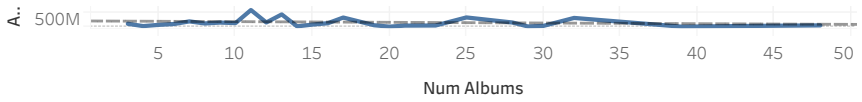
Artist Trends

We also identified some trends regarding the number of albums and the popularity measured in average streams and followers, as well as the p..

Albums-Followers



Albums-Streams



Year-Followers

Year-Streams

This simple analysis showed us that there were relations worth exploring in relation to artist popularity, not only relating to the albums' release, but to the music itself as well.

Datasets and Processes

Considering these findings, we continued with the processing of data in order to create machine learning models that could help predict the pop..

The CSVs that compose our oriignal data, all coming from Billboard charts, and the Spotify API were:

- artistDf
- billboardHot100_1999-2019
- grammyAlbums_1999-2019
- grammySongs_1999-2019
- riaaAlbumCerts_1999-2019..

Which we transformed into:

- artistWeek, by merging artistDf and spotifyWeeklyTop200Streams
- attributesBillboard, by merging songAttributes_1999-2019 and billboard..

To apply the following processes:

- Brainstorming and developing research questions
- Data cleaning and preprocessing
- Model testing
- SQLite integration
- Gathering results..

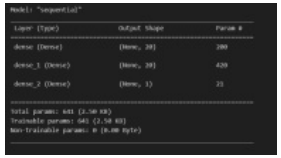
Using tools such as:

- Python
- Pandas
- Numpy
-SQLite..

Models

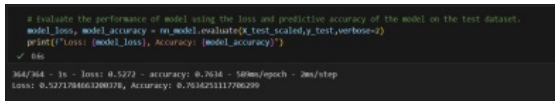
Here are the results from the models we used

01.- Neural Network: Attributes Sample

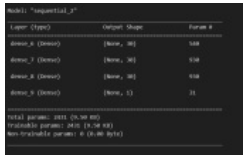


Epochs: 30
Function: Relu/Sigmoid

This model evaluated popularity exclusively agai..

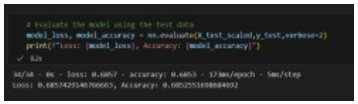


01.- Neural Network: AttributesBillboard



Epochs: 50
Function: Relu/Sigmoid

This model evaluated popularity against song att..

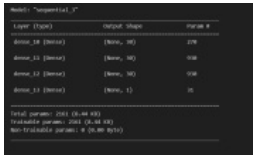


The metrics that we used to evaluate success in all of our models were either popularity or number of followers, and for binary classification purposes, we defined a release as ”is_popular” when it had a value equal or greater than the value at the 75% quartile. Also, the epoch and activation function selection came after iterating different options, and sel..

Models

Here are the results from the models we used

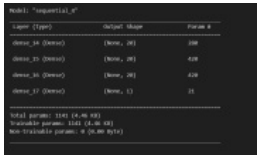
03.- Neural Network: ArtistWeek



Epochs: 50
Function: Relu/Sigmoid

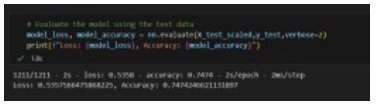
This model evaluated popularity by the amount o..

04.- Neural Network: Attributes Full



Epochs: 30
Function: Relu/Sigmoid

Similar to the first model, this one evaluates pop..

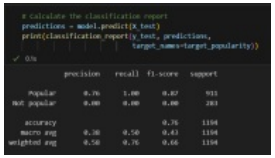


The metrics that we used to evaluate success in all of our models were either popularity or number of followers, and for binary classification purposes, we defined a release as ”is_popular” when it had a value equal or greater than the value at the 75% quartile. Also, the epoch and activation function selection came after iterating different options, and sel..

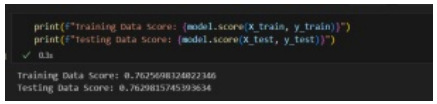
Models

Here are the results from the models we used

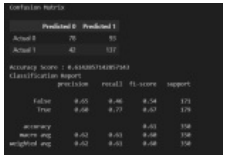
05.- SVC: AttributesBillboard

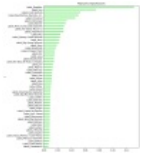


This model evaluated popularity against song attributes, as well as Billboard Charts information, but used an SVC Model instead of Neural Networks.



06.- Random Forest: Grammy/Label





This model, the most different one, didn't evaluate popularity or number of followers, instead it used a Random Forest Model to evaluate the likelihood of getting a grammy in relation to music publishing labels.

The metrics that we used to evaluate success in all of our models were either popularity or number of followers, and for binary classification purposes, we defined a release as "is_popular" when it had a value equal or greater than the value at the 75% quartile. Also, the epoch and activation function selection came after iterating different options, and sel..

Conclusions

- Neural Networks seem to have the highest accuracy scores for these types of evaluations

- Spending considerable effort on cleaning up the data, and performing feature engineering tends to have a much greater impact on model accuracy

- We found that increasing epochs over 50, having more than 3 layers or having more than 30 neurons per layer didn't seem to have a more positive impact on accuracy on any of the tests.

- For our data, it seems that there is a reasonable correlation between musical attributes and success of a musical release, as well as the specifics of the artist, their career history, gender and aggrupation type.

- We have to be very critical in evaluating which features contribute to our target, such as popularity, and which features are consequences of said popularity, and might be impacted by