Proposal: The Curse of the Grammys

DATA 450 Capstone

AUTHOR
Virginia Ferreras

PUBLISHED
February 8, 2024

1 Introduction

Music is a beautiful art that combines sounds and vocals to create forms of expression. Considered a universal language, music has been around for centuries and has evolved to a plethora of different genres. Music artists use music to express not only themselves, but their cultures and creativity. While music is subjective, there are a variety of different ways these musical artists get recognized and rewarded for their works of art. The Grammys is one of the most prestigious awards shows in the music industry, presented by the Recording Academy of the United States. With over 90 different awards, there are many different categories that shine light on the hard work and dedication of the people behind popular bodies of music. The most important awards of the ceremony, titled The "Big Four" are: Album of the Year, Song of the Year, Record of the Year, and Best New Artist.

Of these categories many people wait to see the winner of the Best New Artist award, as it gives upcoming artists a chance to be recognized for their talents. However, this award is also tied to what some may call a curse. Many people speculate that if an artist is awarded this award, then their careers are short lived after, therefore introducing the Best New Artist curse. While the impact these artists may have had on the music industry will still be placed on their respective pedestal, the curse basically points to the decline of their musical achievements, successes, and relevancy after receiving this award. Being an avid music listener and a fan of many different artists, I have always found myself curious about the validity of this curse. In this project, I am going to combine my knowledge of data science with my adoration of music and look into this curse. I want to prove whether or not this curse is true, and if so see if there is a pattern to which artists have fallen off after receiving this award.

2 Dataset

The datasets that I will be using for this project are listed below.

'best_new.csv': This dataset was created using information found on the Grammy's official website. I entered the name of the artists who won the Best New Artist award along with the respective year of the win. There are 43 observations and 2 features found in this dataset, with the date being 1980-2022. The data was found through the Grammy's Official Website.

'chart_data.csv': The dataset was created by utilizing the Billboard's Hot 100's Chart. I went ahead and used web scraping to scrape each weekly chart from the week of February 3, 2024 back to October 12, 1985. There are 200,000 observations in this dataset and 5 features in this dataset. The features include: 'week_of, 'rank', 'song_title', 'artist_name', 'peak_position'. The data was found through the Billboard Official Website.

'spotify_api.csv': This dataset contains several features found through the Spotify API (explained in the next section). There will be 6 features and 43 observations in the dataset. The features include: 'artist_name', 'artist_genre', 'album_name', 'album_year', 'artist_followers', 'artist_popularity'. The data was found through the Spotify for Developers website.

'ama_winners.csv': This dataset will be created by using web scraping techniques to gather the past American Music Award artists who have won an award. There will be 3 features on this dataset that include: 'year_win', 'ama_category', 'artist_name'. The dataset was found through the official American Music Award website.

'bma_winners.csv': This dataset will be created by using web scraping techniques to gather the past Billboard Music Award artists who have won an award. There will be 3 features on this dataset that include: 'year_win', 'bma_category', 'artist_name'. The dataset was found through the official Billboard Music Award website.

*__Variable Explanations:__

- artist_name: Name of the music artist
- week_of: The week of the Billboard Hot 100s chart
- 'rank': The rank of the song on the Billboard Hot 100s chart during that week
- 'song_title': Title of the song on the chart
- 'peak_position': The peak rank that song has had on a Billboard Hot 100's chart.
- 'artist_genre': Genre/s that are associated with the artist
- 'album_name': Name/s of the albums released by the artist
- 'album_year': Year of an albums release
- 'artist_followers': The amount of followers an artist has on Spotify
- 'artist_popularity': The popularity of an artist, which will be indicated by a value between 0 and 100 with 100 being the most popular. This value is calculated from the popularity of the artist tracks. This popularity is calculated by a mathematical algorithm the Spotify API follows, which for the most part is based on the total number of plays a track has and how recent the plays were.
- 'ama_category': American Music Award award category.
- 'bma_category': Billboard Music Award award category.

3 Data Acquisition and Processing

In order to gather the data of the past Grammy Best New Artist winners, I manually entered the artist names and the years they won into an Excel spreadsheet. Once compiled, I converted this spreadsheet into a CSV file.

The Billboard Hot 100s in the music industry's record chart for the United States that is updated weekly. This is a very respected and reputable chart that the music artists and teams use to rank a song's popularity. I want to use these charts to be my main indication of how well an artist does after winning this award. The more singles an artist lands on the charts, the more relevancy and popularity they maintain or grow. To gather data on the Billboard Hot 100s charts, I utilized web scraping techniques. Using the BeautifulSoup library to parse HTML, requests package to make HTTP requests, and the

datetime package to handle dates, I navigated through each year's chart, scraping the necessary data needed. After collecting this data, I utilized the Pandas library to create a dataframe, which was then saved to a CSV file.

While the Grammys is one of the most prestigious music awards in the industry, there are other awards that are as important and well-respected in the music industry. Along with the Grammys, the American Music Awards and the Billboard Music Awards make the"Big Three" major music awards. Therefore, I decided to analyze if these artists have won any awards in these award shows after winning the Best New Artist award. The same scraping method was employed to gather information on past winners of the American Music Awards past winners and Billboard Music Awards.

Spotify is one of the largest and most popular music streaming platforms used by the public. They house a plethora of information regarding different artists on their free API. Therefore, I wanted to utilize this data in order to deepen my future investigations. I accessed the free API system provided by Spotify using my personal Spotify account information to gather more data on the respective artists, including information such as artist genre, album names, release years, followers, and popularity. After compiling the data, I used Pandas to create a dataframe, and then saved the CSV file.

4 Data Processing:

Since the datasets I am working with are already saved as CSV files, I do not have to adjust any of the file formats. After reading my data files using the Pandas file, I will go ahead and begin the process tidying up the data.

1. **Adjusting Column headers:** Since I gathered the data being used for this project, the column headers are already written in a preferred format. However, if needed I will go ahead and adjust column headers to better facilitate future modeling and analysis.

2. **Data Types:** I will check the data types of each dataset and ensure that each feature has the correct data type that will not hinder any future analysis.

3. **Data Encoding:** I will be encoding one variable found in the spotify_api dataset. The genre variable will be investigated during the modeling stage, so I will use one-hot encoding to label each genre as a unique number. This method will help me handle artists associated with a variety of genres.

4. **Data Ranking:** Utilizing the 'rank()' function from the Pandas library, I will be assigning rankings to the 'peak_position' feature, which represents the highest chart position a single has had on the chart (with 1 being the highest).

5. **Handling Missing Values:** The only feature that has missing values is 'peak_position' due to songs that have never charted, thus lacking a peak position. In this case, instead of imputing these missing values with the mean, median, or mode, which would be misleading, I will instead use a placeholder value and impute with the maximum possible value. This preserves the integrity of the data by not assigning artists peak positions that they have never had.

6. **New Datasets:** After tidying the datasets, I will create new data frames that contain only the artists who have won the Best New Artist award from the 'chart_data', 'ama_winners', and 'bma_winners' datasets. This will allow me to focus exclusively on these relevant artists, making future modeling more efficient.

Additionally, I will create a new dataset consisting of the artist name and the count of singles, AMA awards, and BMA awards received. This dataset will be merged with the 'spotify_api' dataset to provide an overview of the artist's popularity, and other relevant information. This will make future visualizations easier to create, and it will place the counts of the awards of an artist in one place.

5 Research Questions and Methodology

I will be investigating three research questions, which are listed below along with the process of how I will answer these questions.

Question 1: Does the success of an artist's career decrease after winning this award? One way I will determine the success of an artist is by looking at the number of singles they have had on the Billboard Hot 100s after winning the Best New Artist award. I will visualize these counts by utilizing Plotly to create an interactive line plot in order to choose which artist you want to see represented on the line chart. The y-axis will be the number of singles, and the x-axis will be the year difference. I will analyze these lines to see if there are any patterns with the number of singles an artist has on the billboard Hot 100s after they win the award.

```
Another way I will determine the success of an artist is by looking at the number of
awards they have received in the American Music Awards and Billboards Music Awards (the
two other important music award shows). I will use a stacked bar plot to visualize the
number of awards these artists received after winning the Best New Artist Award. I will
also use the same line plot that was used to visualize the Hot 100 singles count. This
way, I can look at each artist individually and see the total number of awards they've
one since winning the Best New Artist award.
```

Question 2: Are there any features that affect the success of an artist's career after winning this award? To answer this question, I will use predictive modeling to predict what features have the most effect on an artist's career. A decision tree, random forest, and linear regression will be used to come to a conclusion. I will test its accuracy and ensure that the training and testing data is properly split before starting the modeling process.These models will be able to provide us with a list of the top features that have an effect on the target variable (artist's popularity).

Question 3: Are there differences in the success through an artist's career of the Best New Artist winners across different music genres? In this question, I will be veering away from artists and looking more into a music genres' possible impact on an artist's career. I will create a bar plot that showcases the number of award each genre has won, as well as another bar plot that visualizes the number of singles on the Hot 100s of these genres.

6 Work plan

Week 4 (2/12 - 2/18):

- Scraping the American Music Awards and Billbord Music Awards data (2 hours)
- Gathering Spotify API data (1.5 hours)
- Adjusting/handling column headers, data types, missing values and creating new necessary dataframes (4 hours)

Week 5 (2/19 - 2/25):

- Data encoding and ranking (1 hours)
- Question 1: Line plot (2.5 hours)
- Question 1: Bar plot (1.5 hours)
- Question 2: Start decision tree and random forest models (2 hours)

Week 6 (2/26 - 3/3):

- Question 1: Adjustments (1 hours)
- Question 2: Linear regression (3 hours)
- Question 3: Both bar plots (4 hours)

Week 7 (3/4 - 3/10):

- Question 1, 2, and 3: Adjustments (2 hour)
- Presentation prep and practice (5 hours)

Week 8 (3/11 - 3/17): *Presentations given on Wed-Thu 3/13-3/14. Poster Draft due Friday 3/15 (optional extension till 3/17).*

- Any last minute presentation prep (2 hours)
- Poster prep (5 hours)
- Presentation peer review (1.5 hours)

Week 9 (3/25 - 3/31): *Final Poster due Sunday 3/31.*

- Peer feedback (3.5 hours)
- Poster revisions (3.5 hours)

Week 10 (4/1 - 4/7):

- Blog Post: Draft (3 hours)

- Any changes/adjustments (4 hours)

Week 11 (4/8 - 4/14):

- Blog Post: Draft (4 hours)
- Any changes/adjustments (4 hours)

Week 12 (4/15 - 4/21):

- Final changes and adjustments to code (2 hours)
- Code clean-up/organization (4 hours)
- Blog Post: Draft (2 hours)

Week 13 (4/22 - 4/28): *Blog post draft 1 due Sunday night 4/28.* [All project work should be done by the end of this week. The remaining time will be used for writing up and presenting your results.]

- Draft blog post (4 hours).

Week 14 (4/29 - 5/5):

- Peer feedback (3 hours)
- Blog post revisions (4 hours)

Week 15 (5/6 - 5/12): *Final blog post due Weds 5/8. Blog post read-throughs during final exam slot, Thursday May 9th, 8:00-11:20am.

- Blog post revisions (2 hours)
- Peer feedback (2 hours)

7 References

"Award Nominations and Winners" grammy.com. https://www.grammy.com/awards (accessed Feb. 8, 2024).

"Billboard Hot 100s" billboard.com. https://www.billboard.com/charts/hot-100/ (accessed Feb. 8, 2024).

"Winners Database" theamas.com. https://www.theamas.com/winners-database/ (accessed Feb. 8, 2024).

"Winners Database" billboardmusicawards.com. https://www.billboardmusicawards.com/winners-database/ (accessed Feb. 8, 2024).

"Spotify for Developers" developer.spotify.com. https://developer.spotify.com/ (accessed Feb. 8, 2024).