Billboard Hot 100 Analysis

Laurie Cagney
LCagney@Regis.edu
Summer 2021 – Regis University
MSDS696

Introduction

This project seeks to understand the relationship between Billboard Hot 100 data and Spotify music attributes. This project was completed over the 8-week Summer 2021 term for the Regis University Data Science Program, facilitated by John Koenig. This write-up will detail the research question, dataset, data prep, data model, and conclusions. This project was primarily completed using Python.

Research Question

Can characteristics of a song be used to identify *when* a song was popular on the Billboard Hot 100?

Billboard Hot 100

The Billboard Hot 100 chart is a weekly chart that has been published since 1958. It demonstrates who is popular in music at the current moment. Fans of artists and chart enthusiasts follow these charts to understand where their favorites artists are charting or what the current trends are in music today. Today, the chart position is determined by a combination of sales figures plus streaming figures plus radio impressions. It's an important industry metric that record companies push for. This data will tell me how the general public received the song.

Spotify

Spotify is a streaming service that launched in 2006. Artists are compensated for each stream they receive on a song, paid or free. Spotify is one of the largest streaming services and it offers easy access to their data through an API. This data will tell me how the song sounded.

The Dataset

I combined Spotify data and Billboard Hot 100 data with SQL. I condensed the original dataset to display the final week's entry on the chart alongside its peak position. The dataset contained all the available Spotify characteristics. Prior to data cleaning, I had ~**25K ROWS BY 14 COLUMNS**.

Data Prep

Cleaning the data fell into 3 buckets: what I **removed**, what I **added**, and what I **modified**. This process was partially guided by the output of ProfileReport from pandas_profiling.

Removed	**Added**	**Modified**
High correlated columns	Season	Min/Max scaling rows
Empty rows with no metadata	Isolated Year from Week	Dummy variables
Columns where all data was the same		

When exploring the data, I noticed that it was either all or nothing when it came to having Spotify metadata. Since the song was missing its metadata entirely, and not just part of it, I dropped the whole song rather than try to impute values. I additionally dropped incomplete years from the dataset: 1958 and 2021.

Final dataset was ~**23K ROWS BY 17 COLUMNS**.

In order to proceed with a model, I observed how the data trended over time:

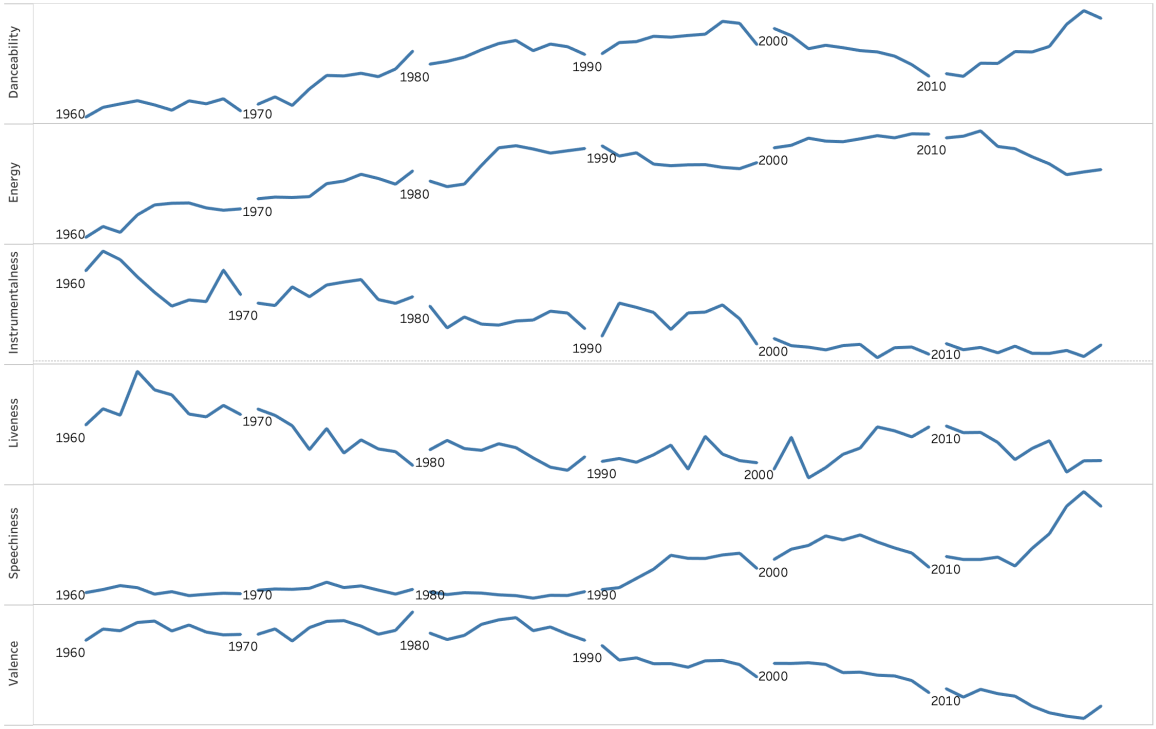

Figure 1: Music characteristics trend over time

The characteristics change on a sliding scale and not all at once as it passes into a new decade. This led me to approach this problem as a **regression** problem instead of classification.

The Data Model

Train/Test Split

I split the data into 80/20 split between train and test. My train dataset had 19k observations while my test dataset had 4,700 observations.

Regression Model

I chose the best model based off the results from LazyPredict: LGBM Regressor. Light Gradient Boosted Machine is a tree-based learning algorithm.

Results

After tuning using GridSearchCV and 10-fold cross validation, the final model had an R^2 value of *77.7%,* RMSE value of *9*, and MAE of *6.4.*

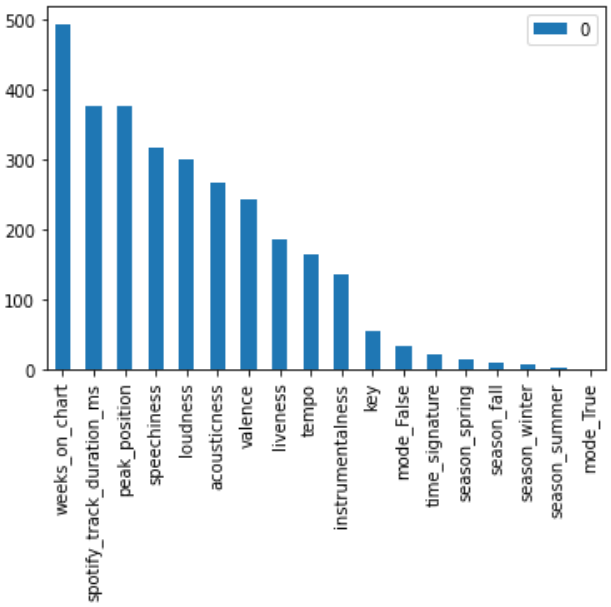

Figure 2: Important Variables

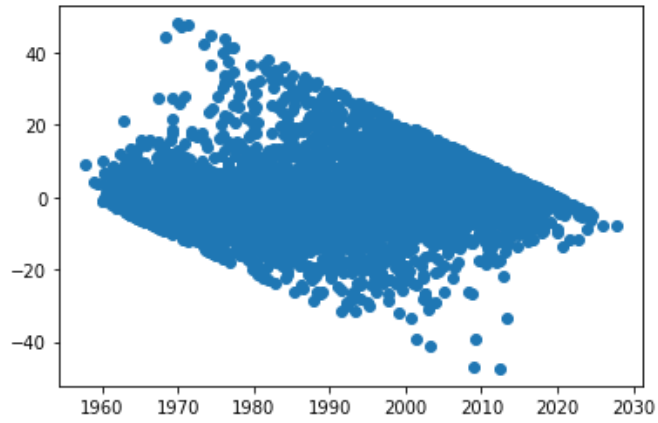

Figure 3: Residual Plot

The important variable plot (Figure 2, left) shows that the added variables I created were not useful to the model. The residual plot (Figure 3, right) showed a linear pattern. This could be due to a lurking variable. My hypothesis is that because I don't know *how* a song charted, whether it was because of its sales, radio impressions, and in recent times, streams, that I'm missing a piece of why it was popular. This also suggests that a song's popularity isn't just because of how it sounds but other factors as well. The other piece that was missing and could improve this is genre or additional Spotify metadata that was not scraped as part of the dataset.

Conclusion

Can we predict it? To a certain extent, yes; however, music sounds and trends are cyclical and often derivative. When digging into the incorrectly predicted song years, I discovered that during 2011 timeframe the TV Show Glee ruled the charts with covers of songs. This threw off the prediction model. Additionally, there are aspects to popularity that may be impossible to quantify. That being said, I believe this model is a good gateway for more improved predictions with additional attributes fed into it.

Recommendations/Next Steps

The following list contains Items I identified that could improve the model or change the direction of the project:

1. Include genre after reducing the cardinality in the data and see how the model can improve its predictions.
    a. Genre was part of the original dataset but it included a list of genres and sub-genres for each song and there was a lot of variability in the label. It would have to be reduced to its main genre and further grouped into smaller broader labels.
2. Tag and remove re-released covers by artists like Glee.
3. Answer the question from a derivation standpoint instead of as "when" – who sounds like who?

Appendix

Additional details

Primary Python Packages Used for Analysis

- Pandas_Profiling
- Matplotlib
- Lazypredict.Supervised
- Sklearn
    - Preprocessing
    - Metrics
    - Model_Selection
- LightGBM

Links

- Data.World Datasets: https://data.world/kcmillersean/billboard-hot-100-1958-2017
- GitHub: https://github.com/lcagney/MSDS-Practicum-2
    - Contains Tableau workbook, code, code references, and cleaned datasets
- YouTube Presentation: https://youtu.be/Q3zYRNRokrc