Introduction

There’s music all around us, but is there a secret algorithm to create the perfect tunes that we can listen to on repeat? We are using a dataset of songs from Spotify, and we want to find the specific features that make songs successful. The dataset gave us data on 18836 songs with different characteristics. Our goal is to create a model that can predict the popularity of a song based on the song's characteristics in combination with other factors.

The music industry is a multi-billion dollar market that is constantly evolving. Although music is an artistic expression, at the end of the day it is also a business. This is what sparked our main question: what characteristics create hit songs? Although there are outside factors such as social media promotion, the timing of release, and trends, we believe that we can also analyze songs algorithmically. To answer this question, we want to look at factors that we can build into a model, such as examining some interesting variables such as energy, liveness, loudness, and more.

Related work

We obtained our data from Kaggle and there was similar work done on it by Mr. Yasser who is the source of the dataset. He uses python for his machine learning models, and he started off with Exploratory Data Analysis where he visualized the features and relationships of all variables. After cleaning and manipulation of data, he created a correlation matrix and tested a regression model. Then he went deeper and used VIF and RFE models and feature elimination using PCA decomposition. Mr. Yasser continued to create multiple regression models and found that polynomial regression predicts better for this data. Other similar work was done by 3 Stanford Computer Science students where they used regression, classification, and backward models. There’s also been a few articles on this topic consisting of similar work.

Data Description

Our data is highly descriptive with over 18,000 data points with 4,000 duplicates that we cleaned up, and we have variables such as song popularity ranking, duration, acoustic, danceability, energy, liveness, loudness, and instrumentals. The popularity ranking is from 1-100 from Spotify, and we can pinpoint the specific variables of these popular songs, and also see the features of songs that are less popular. The more popular songs are ranked 100, with the lesser popular songs being 1. All variables are continuous features, with audio mode, key, and time signature being categorical variables. We created a histogram, and we can see that most of the songs were compiled between 40 to 80 rankings.

Methods and Results

We began modeling by splitting the dataset without duplicates into the training and test dataset. The training dataset is compiled of a random generation of 80 percent of the final dataset. This equates to being around 11 thousand rows. Since we already have the variable called *song_popularity* we used it as our response variable in the logistic regression in order to see which variables have the most significant effect on it. After running the fitting models the *energy, loudness ,danceability, anduio_valence,* and *instrumentalness* ended up having the highest significance on *song_popularity.* The RMSE for the model ended up being 20.36. Since it is pretty high, we wanted to see if there is any other model that can have a better prediction of the *song_popularity.*

Keeping this in mind, we started building the distributed gradient-boosted decision tree. After running the trial decision tree the RMSE was very high, approximately 48. That's why we began tuning the model. After running multiple tuning models, we came to the conclusion that the best possible variables for the max depth and min weight are respectively 5 and 3. The eta used for the final model was 0.005. The subsample and colsample by tree parameters were tuned as well, but didn't have a sufficient effect on the model. Our final model ended up having a lower RMSE of 19 then logistic regression model, but didn't get too low, as we would want it to be.

Moreover, since our model didn't have big success, we decided to focus on determining which variables in the dataset have the effect on the *song_popularity.* The importance matrix determined that i*nstrumentalness, acousticness,* and *loudness* have high importance in the model. To explore this relationship deeper, we ran the shap model, which showed that even though the *instrumentalness* is high on the importance matrix, it is important to notice that it has a negative effect on the *song_popularity*. Whereas, *loudness* and *dancebaility* have a positive effect on it.

The purpose of running two models was to see the difference between RMSE they can get. With our final RMSE being 19, it can be helpful for the music industry to know that with the right amounts of each variable present in the song based on the previous data, it can become very popular or fail to find its listener. This is very important to keep in mind, as the music industry just proved its unpredictability. Moreover, we believe that making a song with less instrumentalism and more loudness and beats will higher the chances of it becoming popular.

Discussion

As previously mentioned, the music and song creation industry is a difficult problem to tackle, but this also creates an opportunity within the machine learning community. The two main insights gained from our models were the significant factors surrounding songs and the specific keys to use. After running our logistic regression model we discovered our significant variables and were able to gain deeper insights through the importance matrix and SHAP value. The largest takeaway that we discovered as a group was how sporadic the SHAP and feature value was across our significant variables. An example of this would be *tempo*, which split equally on both values while also dipping negatively and positively. This created a larger question that pushed us to re-evaluate our framework. Unlike a binary language, music is semantic and it's hard to distinguish. Although you might be thinking why not just classify it as sound? This observation is correct, both music and non-music can exist under this category but from a broader perspective, they are significantly different. To explore this deeper we would need to classify our “sounds/variables” under a higher mathematical alias (such as the classification used in the models of word2vec). The next insight we gained from our models was the significance surrounding musical keys. Once again we had to re-analyze our initial hypothesis. Although we indicated that the third key was to be avoided, without a deeper knowledge of music theory, we once again found ourselves at a crossroads. Musical keys aren't completely distinct due to the minor and major components associated with them. This adds a new dimension to our understanding of the models/exploration results. Two further actions that can be taken off of the back of the project would be to build specific parameters based on genre. This would help narrow sporadic results and focus on more meaningful details. The other action would be to evaluate chords and progressions rather than keys.

Conclusion and Future work

In conclusion, we used a dataset of songs from Spotify to identify the characteristics that make songs successful. We found that the energy, loudness, danceability, audio_valence, and instrumentalness of a song were the most significant factors in predicting its popularity. Our results showed that a distributed gradient-boosted decision tree model with a max depth of 5 and a min weight of 3 had the best prediction for song popularity, with an RMSE of 19. Overall, even though our results weren’t substantial, they helped paint a deeper picture for further analysis.

Contributions

- Anna- modeling, tuning, write up: modeling and results
- Sam - visualizations, model interpretation, problem framing, presentation formatting
- Tanish - data preparation and understanding, tuning

Bibliography (Does not count towards limit) - List of sources used for the project

Data:https://www.kaggle.com/code/yasserh/song-popularity-prediction-best-ml-models/notebook

Tough, David. "Teaching modern production and songwriting techniques: what makes a hit song?." *MEIEA Journal* 13.1 (2013).
https://www.davetough.com/images/ToughMEIEA2013.pdf

Gao, Andrea. "Catching the Earworm: Understanding Streaming Music Popularity Using Machine Learning Models." *E3S Web of Conferences*. Vol. 253. EDP Sciences, 2021.
https://www.e3s-conferences.org/articles/e3sconf/pdf/2021/29/e3sconf_eem2021_03024.pdf