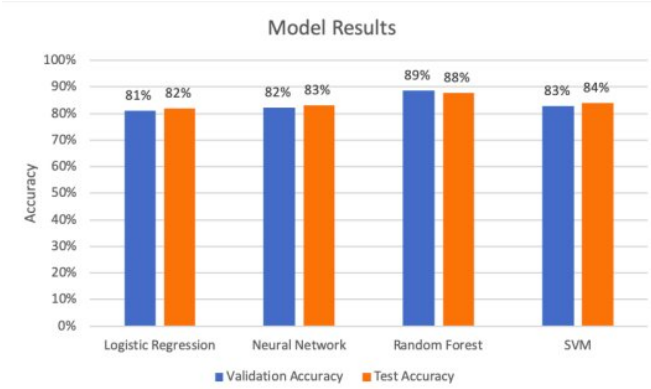
Using Spotify data to predict what songs will be hits

9 September 2019, by Ingrid Fadelli



Model Results on the validation and test sets. Credit: Middlebrook & Sheik.

Two students and researchers at the University of San Francisco (USF) have recently tried to predict billboard hits using machine-learning models. In their study, pre-published on arXiv, they trained four models on song-related data extracted using the Spotify Web API, and then evaluated their performance in predicting what songs would become hits.

"I'm a huge music fan, and I listen to music all day; during my commute, at work, and with friends," Kai Middlebrook, one of the researchers who carried out the study, told TechXplore. "Last spring, I began a research project on automatic music genre classification with professor David Guy Brizan at the University of San Francisco (USF). The project required a large amount of music data, and popular music streaming services have exactly the kind of data I needed."

While he was working on a project related to automatic music genre classification, Middlebrook learned that Spotify allows developers to access its music data. This encouraged him to start experimenting with the Spotify Web API to collect data for his studies. Once he completed the research related to genre classification, however, he set the API aside for some time.

"A few months later, my friend Kian, who is also a data scientist and loves music, and I had a discussion about music," Middlebrook said. "At some point during the conversation, the generally held idea that "all hit songs sound the same" was brought up. We didn't necessarily believe that it was true, but the idea made us wonder: What if hit songs do share some similarities? It seemed possible, so Kian and I decided to investigate further."

Middlebrook and Sheik, who had previously collaborated on the genre classification project, decided to carry out a further investigation using data extracted from Spotify. This new project would also be the final assignment for their data mining course at USF.

"We were collaborating on several other projects for various courses, so it made sense to stick together," Kian Sheik, another researcher involved in the study, told TechXplore. "Lil Nas X's hit "Old Town Road" had just come out of nowhere, and was on the top of the Billboard Hot 100. Kai and I wondered if a computer could have predicted his rise, or if it was just a hit single that came out of left field. What started as a simple final project ended with us exhausting all of state-of-the-art supervised learning models on a large dataset to answer a simple question: Will this song be a hit?"

In their study, Middlebrook and Sheik used the Spotify Web API to collect data for 1.8 million songs, which included features such as a song's tempo, key, valence, etc. They then also collected approximately 30 years worth of data from the Billboard Hot 100 chart.

"Our goal was to see whether hit songs shared similar features, and if so, whether those features

could be used to predict which songs would be hits in the future," Middlebrook said.

The researchers trained and evaluated four different models: a logistic regression, a neural network, a support vector machine (SVM) and a random forest (RF) architecture. During training, these models analyzed a variety of song features, including tempo, key, valence, energy, acousticness, danceability and loudness.

"When given a song, our models would label it with either a one or a zero," Middlebrook explained. "A song labeled with a one means that the model is predicting that the song was a hit. A song labeled with a zero means the model is predicting that the song was not a hit."

The logistic regression model trained by the researchers assumes that song data can be linearly separated into two categories: hits and non-hits. The model assigns a weight to each song feature, and then uses these weights to predict whether a song falls in the "hit" or "non-hit" category.

Logistic regression models have two important advantages: interpretability and speed. In other words, this type of architecture makes it easier to interpret the relationship between explanatory variables (i.e., the song features) and the response variable (i.e., hit or non-hit), and it can also be trained relatively quickly.

The second model trained by the researchers was an RF architecture. This model works by combining a large quantity of building blocks known as decision trees.

"Essentially, a decision tree can be thought of as a model that uses a series of yes/no questions to separate the data," Middlebrook said. "They are interpretable, but prone to overfitting the data. Overfitting means that a model memorizes the training data by fitting it too closely. The problem with overfitting is that the model may not be learning that actual relationship between song features and song popularity because the data often contain irrelevant noise."

To avoid the issue of overfitting, the random forest model used by Middlebrook and Sheik combines hundreds of thousands of decision trees, each of which is trained on a different subset of the training data and a different subset of the song features. The model then makes a prediction (i.e., decides if a song is a hit or non-hit) by averaging the prediction of each tree and combing these results together.

"In our use case, the advantage of the random forest model is its flexibility," Middlebrook said. "It is more flexible than a linear model (e.g. logistic regression)."

The third and fourth models trained by the researchers, namely the SVM and neural network architectures, are both non-linear and are thus harder to interpret. The SVM model works by trying to find the "hyperplane" that best separates the data into the two categories (i.e., hits or non-hits). The neural network architecture, on the other hand, uses one hidden layer with ten filters to learn from the song data.

Among the four models used by Middlebrook and Sheik, the logistic regression model is the easiest to interpret, while the neural network-based one is the hardest. The other two models fall somewhere in the middle.

"Generally, these models will predict based on constraints that they develop through training," Sheik said. "Each model has been trained on the same set of sonic classifiers. The output of the models is tested against historic truth from the Billboard API, whether or not the given track has ever appeared on the Billboard Hot 100 list. We used a fleet of computers at USF to do the number crunching and after a couple weeks of pure computation, we had computed the optimal parameters for each model."

The researchers carried out a series of evaluations to test how well the four models could predict billboard hits. They found that SVM architecture achieved the highest precision rate (99.53 percent), while the random forest model attained the best accuracy rate (88 percent) and recall rate (85.51 percent).

"Recall expresses the ability to find all relevant instances in a dataset, while precision expresses what proportion of data that our model says was relevant actually was relevant," Middlebrook explained. "In other words, recall tell us how likely our model is to accurately predict an actual hit as a hit. Precision tells us the proportion of predicted hits that were actually hits."

According to the researchers, if record labels were to use any of these models to predict what songs will be more successful, they would probably choose a model with a high precision rate than one with a high accuracy rate. This is because a model that attains high precision assumes less risk, as it is less likely to predict that a non-successful song will become a hit.

"Record labels have limited resources," Middlebrook said. "If they pour these resources into a song that the model predicts will be a hit and that song never becomes one, then the label may lose lots of money. So if a record label wants to take a little more risk with the possibility of releasing more hit records, they might choose to use our random forest model. On the other hand, if a record label wants to take on less risk while still releasing some hits, they should use our SVM model."

Middlebrook and Sheik found that predicting a billboard hit based on features of a song's audio is, in fact, possible. In their future research, the researchers plan to investigate other factors that might contribute to song success, such as social media presence, artist experience, and label influence.

"We can imagine a world where record labels who are constantly seeking new talent are inundated with mix-tapes and demos from the "next hot artists,"" Sheik said. "People only have so much time to listen to music with human ears, so "artificial ears," such as our algorithms, can enable record labels to train a model for the type of sound they seek and greatly reduce the number of songs they themselves have to consider."

Classifiers like the ones developed by Middlebrook and Sheik could ultimately help record labels to decide what songs to invest in. Although the idea of using machine learning to skim through demos might be of interest for the music industry, Sheik warns that it could also have undesired consequences.

"While this may be an expedient future, the prospect of a proverbial "chopping block" that artists have to measure up to has the potential to become an echo chamber, or a situation where new music must sound like old music in order to be released on the radio," Sheik said. "Content creators on platforms such as YouTube, which also uses algorithms to decide which videos are shown to the masses, have decried the pitfalls of forcing artists to work for a machine."

According to Sheik, if companies and producers start using algorithms to make artistic decisions, these models should be designed in a way that does not stunt the progress of art. The architectures developed by the two researchers at USF, however, are not yet able to achieve this.

"Novelty bias and other unorthodox features will have to be introduced and invented in order for music as a whole to not approach a cultural singularity at the hands of expedience," Sheik concluded.

More information: Song hit prediction: predicting billboard hits using Spotify data. arXiv:1908.08609 [cs.IR]. arxiv.org/abs/1908.08609

APA citation: Using Spotify data to predict what songs will be hits (2019, September 9) retrieved 16 September 2019 from https://techxplore.com/news/2019-09-spotify-songs.html