IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 12, December 2021

DOI: 10.17148/IJARCCE.2021.101213

Hit Song Predictor by scraping spotify.com, billboard.com and millionsongdataset.com

Kinjal Makwana[1], Priyanka Katore[2],Shrinivas Jawade[3], Tejas Dumane[4], Prof.Poonam Dhamal[5]

G H Raisoni College of Engineering and Management, Wagholi, Pune, Maharashtra (India)[1-5]

Abstract: In this work, we endeavor to take care of the Hit Song Science issue, which plans to foresee which melodies will become diagram besting hits. We develop a dataset with around 1.8 million hit and non-hit tunes and removed their sound elements utilizing the Spotify Web API. We test four models on our dataset. Our best model was arbitrary woods, which had the option to anticipate Billboard melody accomplishment with 88% exactness. In the current review, we moved toward the Hit Song Science issue, planning to anticipate which tunes will become Billboard Hot 100 hits. We grouped a dataset of roughly 4,000 hit and non-hit tunes and extricated every tunes sound highlights from the Spotify Web API. We had the option to anticipate the Billboard accomplishment of a tune with around 75% precision on the approval set, utilizing four AI calculations. The best calculations were Support vector machine, Logistic Regression and a Deep learning.

Keywords: Machine Learning ,Hit Song Science ,Classification ,Data Mining, Data Collection

INTRODUCTION

The Billboard Hot 100 Chart stays one of the conclusive ways of estimating the achievement of a well known melody. We examined utilizing AI methods to anticipate whether or not a melody will turn into a Billboard Hot 100 hit, in light of its sound highlights. The contribution to every calculation is a progression of sound highlights of a track. We utilize the calculation to yield a paired expectation of whether or not the melody will highlight on the Billboard Hot 100. This examination is pertinent to artists and music marks. Not exclusively will it assist with deciding how best to deliver tunes to augment their potential for turning into a hit, it could likewise assist with concluding which tunes could give the best return for speculation on promoting and exposure. Besides, it would help craftsmen and music names figure out which melodies are probably not going to become Billboard Hot 100 hits.

I. METHODS

Dataset and Features: A dataset of 10,000 irregular tunes was gathered from the Million Songs Dataset (MSD) , a free dataset kept up with by labROSA at Columbia University and EchoNest. This was reduced to tunes delivered somewhere in the range of 1990 and 2018. Then, we gathered a dataset of all remarkable tunes that were included on the Billboard Hot 100 between 1990-2018, utilizing the Billboard API library . The datasets gave the craftsman name and melody title, just as other different highlights. To adjust the dataset between certain (hits) and negative (non-hits) models, we eliminated 66% of the melodies gathered from the Billboard Hot 100. At last, we eliminated covering tunes to frame a dataset of around 4,000 melodies. Tracks were named 1 or 0: 1 showing that the melody was included in the Billboard Hot 100 (between 1991-2010) and 0 demonstrating in any case. Then, we utilized the Spotify API to extricate sound highlights for these melodies . The Spotify API gives clients 13 sound highlights, of which we picked nine for our investigation: Danceability, Energy, Speechiness, Acousticness, Instrumentalness, Liveness, Valence, Loudness, and Tempo. The initial seven highlights are addressed as qualities somewhere in the range of 0 and 1 by Spotify. Tumult is estimated in decibels and rhythm alludes to the speed of the melody in beats each moment. To represent craftsman recognisability, we characterized an extra measurement: the craftsman score. Every tune was alloted a craftsman score of 1 assuming the craftsman had a past Billboard Hot 100 hit, and 0 in any case. We thought back to 1986 for this measurement. There is some intrinsic mistake in this action. Assuming that a craftsman had a hit melody before 1986, yet not later, they were given a craftsman score of 0.

II. ALGORITHM

To anticipate a song0 s achievement, we utilized six distinctive AI calculations:

Expectation Maximization (EM), Logistic Regression (LR), Gaussian Discriminant Analysis (GDA), Support Vector Machines (SVM), Decision Trees (DT), and Neural Networks (NN). We zeroed in for the most part on the exactness of results, yet we report the accuracy and review also. Bogus positive expectations might

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 12, December 2021

DOI: 10.17148/IJARCCE.2021.101213

be expensive assuming a music mark puts resources into a tune that is probably not going to turn into a hit. For an underlying recognizable proof of groups in the information, we utilized the EM calculation expecting no marked information, then, at that point, contrasted the bunches with the genuine names. This calculation makes bunches of the information, as indicated by a predefined likelihood conveyance. In every cycle, the boundaries of each bunch are determined, and the likelihood of every information point being in each group is determined.

To anticipate whether or not a melody will be a Billboard hit, we utilize four distinct models:

- Logistic Regression (LR)
- Neural Network (NN)
- Random Forest (RF)
- Support Vector Machine (SVM)

Logistic regression is a famous arrangement calculation. It is utilized when the reliant (target) variable is downright. The thought in LR is to track down a connection among highlights and the likelihood of a specific result There are two sorts LR issues twofold calculated relapse and multi-class strategic relapse. We utilized twofold calculated relapse on the grounds that our reliant variable has two potential qualities 0 (non-hit) and 1 (hit). We utilize the sigmoid enactment capacity to compel our likelihood gauge somewhere in the range of 0 and 1.

$$\sigma(\mathrm{x}) = \frac{e^{x}}{1+e^{x}}$$

We utilize Maximum Likelihood Estimation (MLE) to appraise the component coefficients and RMSEprop to back-engender the inclinations north of 1000 ages. We characterize the expense work beneath

$$\mathrm{L}(\beta;\ \mathrm{y}) = \prod_{i=1}^{n} \mathrm{P}(\mathrm{Yi} = yi|\mathrm{Xi} = xi)$$

$$= \prod_{i=1}^{n} \sigma\,(x_i^t\ \beta)^{yi}(1-\sigma(x_i^t\ \beta))^{1-yi}$$

Where $\sigma(x_i^t\ \beta)$ is the probability of a hit and $(1-\sigma(x_i^t\ \beta)$ is the probability of a non-hit. Additionally, yi = 1 (hit) or 0 (non-hit).

Neural Networks (NNs) have become well known to address arrangement assignments later the ascent of profound learning. We utilize a straightforward neural organization with one secret layer to address HSS. We use RMSprop-an unpublished improvement calculation intended for neural organizations, first proposed by Geoff Hinton, and sigmoid capacity in the last layer to compel the result somewhere in the range of 0 and 1. In the secret layer, we utilize ten channels and corrected direct unit (ReLU) initiation. We set the clump size to 32 and quit preparing later 1000 ages.

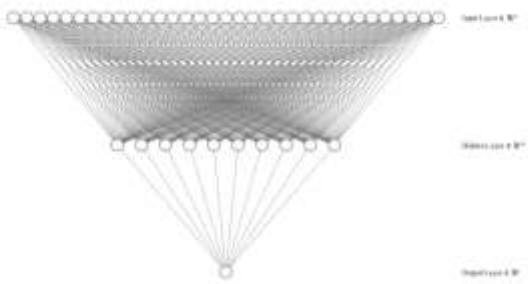

Diagram of our neural network architecture. We use 1 hidden layer with 10 filters.

Random Forest (RF) models are quite possibly the most famous gathering method utilized in order. These models expect to address for the issue of over-fitting in conventional choice trees. This won't be shrouded top to bottom, however choice trees will quite often learn on sporadic ways of information. RF models train various profound choice trees on various parts of the dataset determined to diminish the general change. In addition to the fact that RF was the most reliable model generally speaking, yet it was the speediest to prepare. We utilized a greatest number of highlights of eight with 80 assessors and a base split state of two examples under the Gini model.

Support Vector Machine (SVM) intends to observe the most ideal hyper-plane that isolates the information into two particular classes. We utilized the Gaussian Radial Basis Function (RBF) as our portion: $\exp(-\gamma||\mathrm{x}-\mathrm{x}'||^2)$.Our model uses $\gamma = 0.1$ and $C = 10$

III.RESULT

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 12, December 2021

DOI: 10.17148/IJARCCE.2021.101213

We zeroed in chiefly on the exactness of results, yet we report the accuracy and review also since bogus positive expectations might be expensive when a music mark puts resources into a melody that is in reality improbable to turn into a hit (Table 1).

Table 1: Model Results

Models	Accuracy		Precision		Recall	
	Test	Val	Test	Val	Test	Val
Logistic Regression	0.8151	0.8065	0.7526	0.7457	0.9391	0.9298
Neural Network	0.8214	0.8305	0.8235	0.8233	0.7913	0.7671
Random Forest	**0.877**	**0.887**	**0.86**	**0.87**	**0.9**	**0.89**
SVM	0.839	0.828	0.995	0.993	0.704	0.706

The NN model with one secret layer gave 82.14% and 83.05% exactness on the approval and test information, with comparable outcomes on the preparation information demonstrating no over-fitting. The last cross-entropy misfortune later 1000 ages was 0.4261. The accuracy and review on the approval set were 82.33% and 76.71%. The disarray grid on the approval set shows that there are some bogus up-sides (Table 2).

Table 2: NN Confusion Matrix on the validation set

		Actual	
		Hit	Non-Hit
Predicted	Hit	1027	363
	Non-Hit	42	707

The LR model yielded 80.65% precision on the approval information and 81.51% exactness on the test information, with comparable outcome on the preparation information showing no over-fitting. The accuracy and review were OK at 74.57% and 92.98%. The disarray grid on the approval set shows that there are some bogus up-sides (Table 3).

Table 3: LR Confusion Matrix on the validation set

		Actual	
		Hit	Non-Hit
Predicted	Hit	994	339
	Non-Hit	75	731

The RF model yielded 88.7% exactness on the approval information and 87.7% precision on the test information, with comparable outcome on the preparation information demonstrating no over-fitting. The accuracy and review were satisfactory at 87% and 89%. The disarray framework on the approval set shows that there are some bogus up-sides and bogus negatives (Table 4).

Table 4: RF Confusion Matrix on the validation set

		Actual	
		Hit	Non-Hit
Predicted	Hit	917	153
	Non-Hit	82	993

The SVM model yielded 82.8% precision on the approval information and 83.9% exactness on the test information, with comparative outcome on the preparation information demonstrating no over-fitting. The accuracy and review were OK at 79% and 89%. The disarray grid on the approval set shows that there are some bogus negatives (Table 5).

Table 5: SVM Confusion Matrix on the validation set

		Actual	
		Hit	Non-Hit
Predicted	Hit	1065	5
	Non-Hit	447	628

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 12, December 2021

DOI: 10.17148/IJARCCE.2021.101213

IV. CONCLUSION AND FUTURE WORK

The outcomes showed that SVM and RF beat LR and NN concerning exactness (Figure 1). The most vigorous model is the RF. Curiously; the SVM had the most elevated accuracy exactness (Table 1). The bogus positive rate for our SVM is exceptionally low, while keeping a normal bogus negative rate. Reality esteems anticipated by this model can be trusted while the bogus qualities can't. This calculation is voracious and will accept minimal measure of hazard while characterizing a positive. Music names might like to utilize the SVM since it is doubtful to foresee hits erroneously. In later examinations we might want to explore name impact and online media presence as for melody achievement. Utilizing highlights of the actual sound joined with craftsman past-execution has figured out how to clarify a greater part of the difference in the information; we accept there are more sorts of elements which can furnish our model with a social setting to make stunningly better expectations.

The investigation showed that LR and NN yielded the most elevated exactness, accuracy and review of the calculations tried. SVM and DT experienced over fitting. We might want to utilize more information to lessen the fluctuation of results. Rather than utilizing 4,000 tunes, we desire to incorporate all Billboard Hot 100 hits taken from a more extended time span, and a comparable number of non-hits from the MSD. Besides, we might want to investigate extra sound highlights, for example, span, which was excluded from this undertaking yet can possibly anticipate a melodies Billboard achievement.

V. ACKNOWLEDGEMENT

I would like to thanks to Mrs. Poonam Gupta for her guidance and support. I also thankful of department of Information Technology and G H Raisoni College of Engineering and Management Wagholi, Pune for providing essential research facilities.

VI. REFERENCES

[1] Billboard. (2018). Billboard Hot 100 Chart. Retrieved from: https://www.billboard.com/charts/hot-100
[2] Chinoy, S. and Ma, J. (2018). Why Songs of the Summer Sound the Same. Nytimes.com. Retrieved from: https:// www.nytimes.com/interactive/2018/08/09/opinion/dosongs- of-the-summer-sound-the-same.html
[3] Guo, A. Python API for Billboard Data. Github.com. Retrieved from: https://pypi.org/project/billboard.py/
[4] Dorien Herremans, David Martens Kenneth Srensen (2014) Dance Hit Song Prediction, Journal of New Music Research, 43:3, 291-302.
[5] Mauch, M., MacCallum, R. M., Levy, M., and Leroi, A. M. (2015). The Evolution of Popular Music: USA 19602010. R. Soc. open sci.
[6] Sander Dieleman and Benjamin Schrauwen, Multiscale approaches to music audio feature learning, in Proc. Int. Soc. Music Information Retrieval Conf., 2013, pp. 116121.
[7] Singhi, Abhishek, and Daniel G. Brown. "Hit song detection using lyric features alone." Proceedings of International Society for Music Information Retrieval (2014).
[8] Spotify Web API. Retrieved from: https://developer.spotify.com/
[9] Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. The Million Song Dataset. In Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011), 2011.
[10] Yu, Lang-Chi, et al. "Hit Song Prediction for Pop Music by Siamese CNN with Ranking Loss." arXiv preprint arXiv:1710.10814 (2017).

