BINARY CLASSIFICATION WITH SPOTIFY SONGS: HOW DOES SPOTIFY RECOMMEND SONGS?

Have you ever wondered how Spotify knew what songs you might like or dislike? If so, how did they do it? What type of features did they use to predict if a user would like a certain song or not? Which features are most important? These are all imperative questions in trying to accurately classify unseen Spotify songs based on a user's music taste.

AUTHOR

Sedrick Thomas - Vice President of RIT AI,
3rd Year Management Information Systems Major, and
minor in Software Engineering

AFFILIATIONS

RIT AI Club

Introduction

I conducted this research because I was curious if I could use Machine Learning to understand my Spotify listening patterns. This led me to develop a program to acquire my listening data using the Spotify API. I was able to acquire 1,000 songs worth of labeled data. 500 liked songs and 500 disliked songs. Each song contains 13 features that describe the content of the song itself. Some of these features include energy, tempo, mode, danceability, etc.

Objective

My goal is to build a prediction model that takes in this data and classifies unseen songs with great accuracy (> 70%).

Methodology

Conducted Exploratory Data Analysis on the full dataset. Then, I split the data into training and testing, 80 and 20% respectively. Next, I created a pipeline to scale down numerical values and one-hot-encode categorical values. I used these transformed values to train 5 different models and find the best performing model.

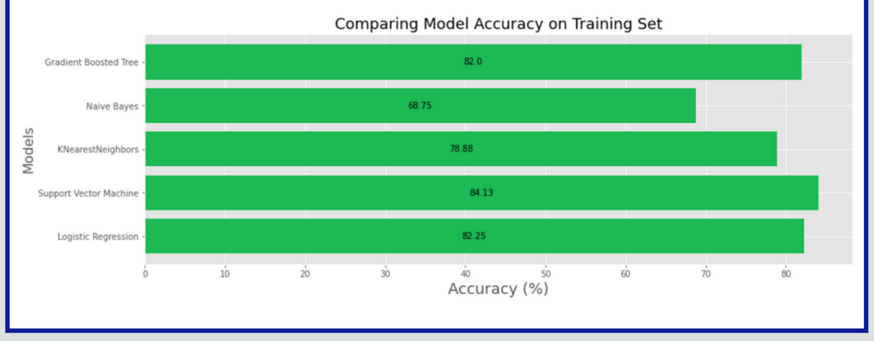
Results

The top 3 performing models on the training set were the Support Vector Machine (84.13%), Logistic Regression (82.25%), and Gradient Boosted Tree (82%). I decided to hyperparameter tune the Gradient Boosted Tree and Support Vector Machine and evaluate it on the test set. The Support Vector Machine had 82% accuracy while the Gradient Boosted Tree had 77.5% accuracy.

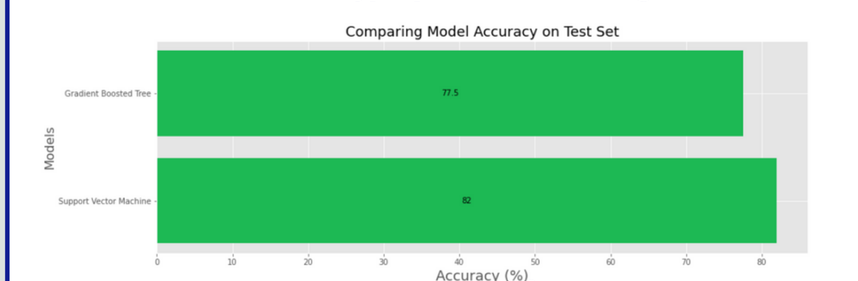
Analysis

- The Support Vector Machine performed the best out of all 5 prediction models with an accuracy of 82%.
- Basically, It can predict 8 out of 10 unseen songs correctly on the test set.
- When training the Gradient Boosted Tree, the most important features were loudness and speechiness.
- Loudness (0.40) and Energy (0.39) had a weak to moderate correlation with being classified as a liked song.

Before Hyperparameter Tuning



After Hyperparameter Tuning



Conclusion

- Although the model has a relatively high accuracy on the test set, it isn't perfect.
- For instance, there may be some false negatives on unlabeled data where I actually like the song and vice versa.
- Ways to mitigate this is to gather more data. More specifically, I should gather more disliked songs.
  - This will allow the model to better differentiate between a disliked song and liked song.

Predictions

	name	artist	Actual	Prediction
521	Rock Wit'cha	Bobby Brown	1	1
737	Pilgrimage	Ichiko Aoba	1	1
740	See Me	Rich Brian	1	1
660	Memories	Levox	1	1
411	Old School	Toby Keith	0	0
678	Cold As You	Luke Combs	0	0
626	Weatherman	Wild Rivers	0	0
513	The Good Ones	Gabby Barrett	0	0
859	Recent Thoughts (Remastered)	LANO	1	1
136	Battling	Konola	0	1
811	I Was Chasing a Firefly to Light My Way Home b...	stream_error	1	1
76	Who Cares	Filmore	0	0
636	Same Truck	Scotty McCreery	0	0
973	See You Tomorrow	Muscadine Bloodline	0	0
938	I Can't (feat. Old Dominion)	Caitlyn Smith, Old Dominion	0	0
899	Now Or Never - Bonus Track	Kendrick Lamar, Mary J. Blige	1	1

Features

	id	name	artist	danceability	energy	key	loudness	mode	speechiness	acousticness
0	6olahjoSHQ3GQJg0hrLdzY	I Believed It (feat. Mac Miller)	dvsn, Ty Dolla $ign, Mac Miller	0.524	0.699	1	-5.559	0	0.0462	0.3030
1	2u93OtblDZphLnmRFjBeeg	Color of Autumn	Nujabes	0.701	0.602	1	-12.291	0	0.0557	0.6350
2	3HfJucp1eH2WZYn9fdDE74	c l o s e	J. Cole	0.724	0.912	9	-5.628	1	0.4040	0.3130
3	73aofsDvxQznhWLfgQ5gBz	Grew Apart	Logan Mize, Donovan Woods	0.624	0.640	7	-7.735	1	0.0806	0.0529
4	07Jhq6iNvTWXFShXK3kXHy	All I Know About Girls	Old Dominion	0.683	0.500	4	-8.331	1	0.0275	0.5960

RELATED LITERATURE

Saravanou, A., Tomasi, F., Mehrotra, R., & Lalmas, M. (2021, October 29). Multi-task learning of graph-based inductive representations of music content. Spotify Research. Retrieved September 25, 2022, from https://research.atspotify.com/multi-task-learning-of-graph-based-inductive-representations-of-music-content/

Web API reference: Spotify for developers. Home. (n.d.). Retrieved September 25, 2022, from https://developer.spotify.com/documentation/web-api/reference/#/operations/get-audio-features