Group Project: Music prediction

ISM 6136 Data Mining

Dr. Mohammadreza Ebrahimi

Nov 14th, 2021

University of South Florida, Tampa.

GROUP 5: -

Ankit Reddy Baddam

Shubodai Reddy Chalamalla

Srikar Reddy Gundannagari

Venkata Sai Thrinesh Munagapati

Summary of Contents

Summary of Figures

Background Of the Problem:

The online music streaming services allow users to play and listen to the music of their choice anytime and anywhere. This is facilitated by the wide-scale penetration of the internet and the increase in users of digital devices. Furthermore, the exponential growth rate of the global population, rapid urbanization, rise in disposable income, and love for music among the population foster the growth of the global online music streaming market in the forthcoming years.

With over 10 global music streaming platforms and many more streaming platforms regionally this market size is projected to reach US$23,293m in 2021. Revenue is expected to show an annual growth rate (CAGR 2021-2025) of 10.08%, resulting in a projected market volume of $34,198m by 2025. In the Music Streaming segment, the number of users is expected to amount to 913.2m users by 2025. User penetration will be 8.4% in 2021 and is expected to hit 11.7% by 2025.

The constantly growing number of subscribers on the music streaming platforms is increasing the audience globally. High-quality, free trials, and discounted prices offered by music streaming service providers are attracting a huge customer base. According to the International Federation of Phonographic Industry (IFPI), around 62.1% of the revenue of the global recorded music market is generated from online streaming. These factors are responsible for propelling the growth of the global online music streaming market. The number of users directly affects the revenue generated or more precisely the profit or loss. Spotify a global music streaming service that was started in 2006 with 104mil and revenue of €2.93 billion has exponentially grown to have a market cap of €7.85 billion with 365 million users.

Users can stream varied songs globally means that the streaming platforms are besieged by huge data. Sorting out all this digital music is very time-consuming and causes information overload. As the number of choices is immense, there is a need to filter, prioritize and efficiently deliver relevant information to diminish the problem of information overload, which has been a potential problem for many users. Many streaming platforms have understood this business problem and have concentrated on solving this using the recommendation system to retain the customer base giving an enhanced personalized experience.

Motivation for solving the problem

"*Acquiring a new customer can cost five times more than retaining an existing customer*" increasing customer retention by 5% can increase profits from 25-85%. The success rate of selling to a customer you already have is 60-70%, while the success rate of selling to a new customer is 5-20%. The figures stated above show clearly, how important it is a task to retain their customer base with the immense competition from other platforms offering the same or more advanced features with nearly similar subscription costs.

Music recommender systems have been proven to be a beneficial means for online users to cope with information overload and have become one of the most powerful and popular tools. It is software that is used to analyze users' tastes and provide them with a list of music that they would like to listen to. It helps to analyze users' preferences and provide them with lists of

products that they would like to prefer. This project is an investigation of using collaborative filtering techniques for a music recommender system. Collaborative filtering is the technology that focuses on the relationships between users and between items to make a prediction. The goal of the music recommendation system is to compute similarities between genres of the songs played by the users and generate the playlist accordingly. This will enhance user personalization and improve the overall user experience, thus aiding in customer retention.

Solution Methodology

To achieve our goal, we are using the following dataset provided by Spotify's API from Kaggle. The dataset has around 17 attributes which have been filtered to 16 attributes. For creating the evaluation metrics, we would be relying on the basic parameters like ROC to understand the distribution of true positive rate and false-positive rate. Additionally, we would be using general algorithmic metrics which help us understand which algorithm works better like Precision, Accuracy, Recall, and F1 Score.

We have implemented the solution using 3 algorithms 2-class boosted decision tree, 2-class neural network, and 2-class logistic network. To improve the solution, we would be using a Partition sample, under which we would be using stratified sampling and feature-based methodology to evaluate feature selection and improve the robustness of the model. Furthermore, we tuned the model for each algorithm by adjusting the parameters. Another important aspect of our solution would be changing one parameter at a time for understanding how each parameter was working under the hood to improvise the training performance. The next aspect we have looked at in our evaluation metrics is the AUC-ROC curve, which plots curve plots for both independent variables and dependent variables at different classification thresholds. Lowering the classification threshold classifies more items as positive, thus increasing both False Positives and True Positives. Preferable models cover a greater area under the curve, which results in a higher overall rate of true positives and false positives. It was observed that the neural network was covering more area when compared to the boosted decision tree and logistic regression model. So, using this additional criterion we were able to decide which algorithm was working better.

The Confusion Matrix is another commonly used evaluation metric for classification problems. A good model will have high True Positives and True Negatives but low False positives and False Negatives. This metric would also help us understand which algorithm would be working better when compared to the other. For this case, our main objective was to set the number of False positives to the lowest and wanted True Negatives to be on a median so that the user would be recommended songs different from his preference. This in fact can improve our recommendation model as we are limited in predicting the current mood of the user, this would let the user like or dislike. Either way, it would enhance the prediction model and give the user an experience that their interests are considered.

Description of the dataset:

The dataset consists of 2017 instances and 16 attributes. The attributes in the dataset are as follows: 'Acousticness', 'danceability', 'duration_ms', 'energy', 'instrumental Ness', 'key', 'liveness', 'loudness', 'mode', 'speechiness', 'tempo', 'time signature', 'valence', 'target', 'song title', 'artist'. All the attributes are explained in detail below:

- **Acousticness**: It is music that solely or primarily uses instruments that produce sound through acoustic means. If the speechiness of a song is above 0.66, it is probably made of spoken words, if it is between 0.33 and 0.66 then it is a song that may contain both music and words and further the score below 0.33 indicates, the song does not contain any speech.
- **Danceability**: This attribute is calculated using a mixture of song features such as beat strength, tempo stability, and overall tempo. The value obtained determines the ease with which a person could dance to a song over the period of the entire song.
- **Duration_ms:** Duration is the length of the total playtime of the song.
- **Song energy**: It is the sense of forwarding motion in music, which keeps the listener engaged and listening.
- **Instrumentalness**: This Value represents the number of vocals in the song. If the score is closer to 1.0, the song is more instrumental. A score of 1.0 indicates that the song likely to be an acoustic one.
- **Key:** The key to a song is the note or a chord the music is centered around the tonic.
- **Liveness:** It refers directly to the time required for the sound to decay in a room.
- **Loudness**: Loudness is a way to measure audio levels based on the way humans perceive sound. In our case, it is being measured in negative decibels. The negative decibels measure how much quitter something is then the threshold of human hearing.
- **Mode**: A mode is the vocabulary of a melody. It specifies which notes can be used and indicates which have special importance.
- **Speechiness:** It indicates the presence of spoken words in a track. The more exclusively speech-like the recording, the value is closer to 1.0.
- **Tempo:** It measures how fast or slow a piece of music is performed. Tempo is generally measured as the number of beats per minute. 60 bpm implies one beat will be played each second. The higher the value, the faster they play.
- **Time Signature:** This attribute indicates how many counts are present in each measure and which type of note will receive one count.
- **Valence:** This shows the musical positiveness conveyed by a track. It ranges from 0.0 to 1.0 describing the musical positiveness conveyed by a track.
- **Song Title:** It refers to the name of the song.
- **Artist:** An artist is someone who composes or performs or releases a song
- **Target:** Target is the dependent variable and is classified as 1 or 0.

  1 – Likes the song

  0 – Dislikes the song

Experiment created on 11/4/2021 › data.csv › dataset

rows 2017 columns 17

Column 0	acousticness	danceability	duration_ms	energy	instrumentalness	key	liveness	loudness	mode	speechiness	tempo	time_signature	valence	target	song_title	artist
0	0.0102	0.833	204600	0.434	0.0219	2	0.165	-8.795	1	0.431	150.062	4	0.286	1	Mask Off	Future
1	0.199	0.743	326933	0.359	0.00611	1	0.137	-10.401	1	0.0794	160.083	4	0.588	1	Redbone	Childish Gambino
2	0.0344	0.838	185707	0.412	0.000234	2	0.159	-7.148	1	0.289	75.044	4	0.173	1	Xanny Family	Future
3	0.604	0.494	199413	0.338	0.51	5	0.0922	-15.236	1	0.0261	86.468	4	0.23	1	Master Of None	Beach House
4	0.18	0.678	392893	0.561	0.512	5	0.439	-11.648	0	0.0694	174.004	4	0.904	1	Parallel Lines	Junior Boys
5	0.00479	0.804	251333	0.56	0	8	0.164	-6.682	1	0.185	85.023	4	0.264	1	Sneakin'	Drake
6	0.0145	0.739	241400	0.472	0.000007	1	0.207	-11.204	1	0.156	80.03	4	0.308	1	Childs Play	Drake
7	0.0202	0.266	349667	0.348	0.664	10	0.16	-11.609	0	0.0371	144.154	4	0.393	1	Gyöngyhajú lány	Omega
8	0.0481	0.603	202853	0.944	0	11	0.342	-3.626	0	0.347	130.035	4	0.398	1	I've Seen Footage	Death Grips
9	0.00208	0.836	226840	0.603	0	7	0.571	-7.792	1	0.237	99.994	4	0.386	1	Digital Animal	Honey Claws
10	0.0572	0.525	358187	0.855	0.0143	5	0.649	-7.372	0	0.0548	111.951	3	0.524	1	Subways - In Flagranti Extended Edit	The Avalanches
11	0.0915	0.753	324880	0.748	0.00348	10	0.212	-8.62	1	0.0494	104.322	4	0.642	1	Donme Dolap - Baris K Edit	Modern Folk Üçlüsü
12	0.253	0.603	356973	0.434	0.0619	0	0.108	-11.062	1	0.0342	127.681	4	0.381	1	Cemalim	Erkin Koray
13	0.366	0.762	243270	0.476	0	0	0.103	-12.686	1	0.114	130.007	4	0.367	1	One Night	Lil Yachty
14	0.44	0.662	247288	0.603	0	9	0.0972	-8.317	0	0.0793	125.011	4	0.351	1	Oh lala	PNL
15	0.019	0.637	188333	0.832	0.0563	6	0.316	-6.637	1	0.163	99.988	4	0.317	1	Char	Crystal Castles
16	0.0239	0.603	270827	0.955	0.0451	1	0.119	-4.111	1	0.0458	123.922	4	0.773	1	World In Motion	New Order

Figure 1 – Dataset Visualization

Algorithm Comparison:

The first approach we considered was the Two-Class Boosted Decision Tree because we were working with a binary classification problem. It uses an ensemble approach to improvise its performance. In this, the second tree corrects the first tree's errors, and the third tree corrects the first and second tree's errors, and so forth. Initial accuracy was 76.2 percent at the 0.43 threshold, and accuracy increased to 76.4 percent at the 0.32 threshold after tuning the model by increasing the number of leaves per tree from 20 to 30 and the learning rate from 0.2 to 0.3. On further increasing the number of leaves of a tree, the precision is increased but the accuracy is decreased. This might be the case of overfitting.

The next algorithm we considered for this experiment was two-class logistic regression. We initially achieved an accuracy of 68% with L1 and L2 values as "1" after running the logistic regression model. We tried tuning the model to improve performance. Regularizing is done to simplify the model and prevent overfitting. We tried running the model by lowering the L1 and L2 regularization weights close to zero and at 0.005, the accuracy increased by 10.2%, which was a significant improvement over the prior performance and the final accuracy we achieved was 78.2%.

The next algorithm we considered for the experiment was a Two-class Neural Network, which is a set of interconnected layers. Input and output layers are connected with hidden layers comprised of weight edges and nodes. There is a lot to play with the parameters in this model. Experiments were run using 100,110,130,150,180 and 200 as no of hidden nodes. Initial accuracy was 78.9% at 100 hidden nodes and 0.1 learning rate. At 110 hidden nodes, and a learning rate of 0.5 and upon decreasing the initial learning weight to 0.005 the accuracy had increased by 3% to 81.9% at a threshold of 0.56.

Summary of Results:

By studying the data, the attributes acousticness, danceability, valance, instrumentalness, tempo, and artist name were known to be the most important features in predicting the type of music. It was very interesting to know that by using the L1, L2 regularization by keeping the values close to 0, we were able to avoid the overfitting in Logistic regression. In addition to this, it also made the model simple. This indicates the importance of data transformation.

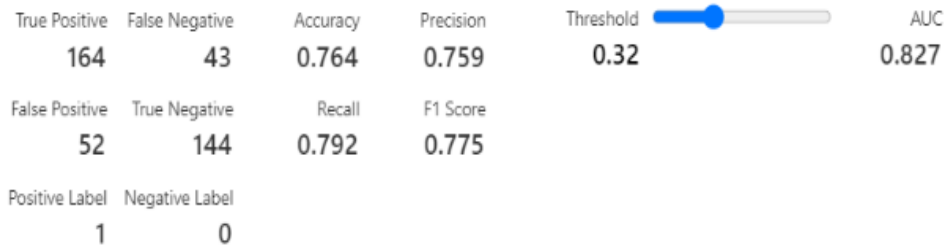

Figure 2 – Two-class Boosted Decision Tree Metrics

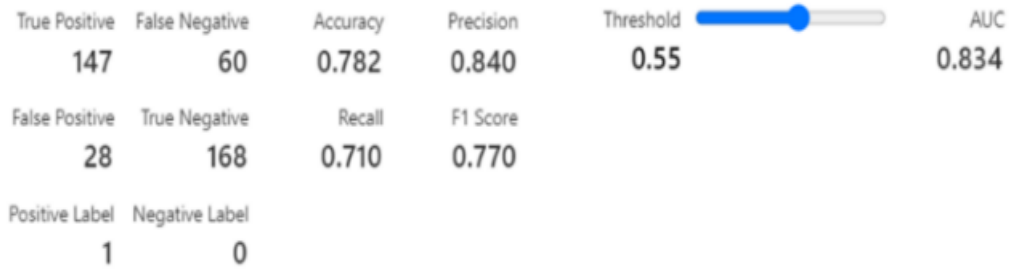

Figure 3 – Two-class Logistic Regression Metrics

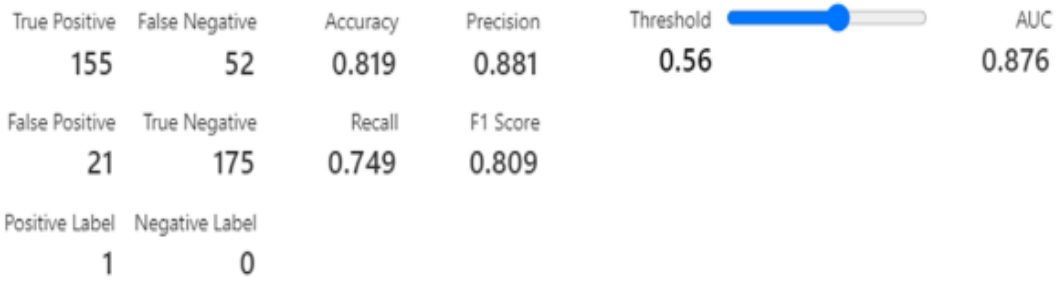

Figure 4 – Two-class Neural Network Metrics

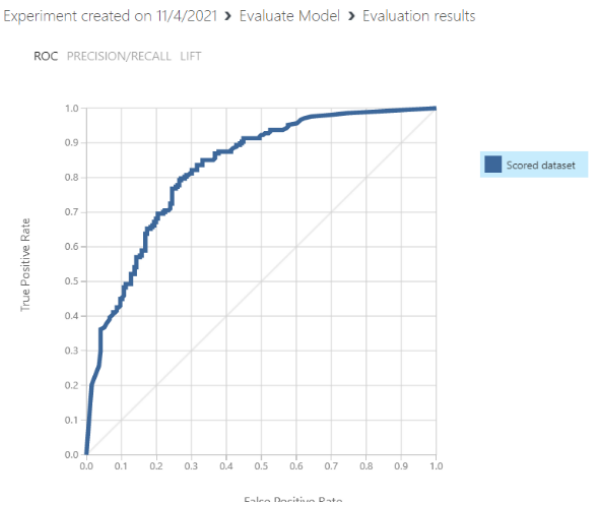

Figure 5 – Two-class Boosted Decision Tree ROC curve

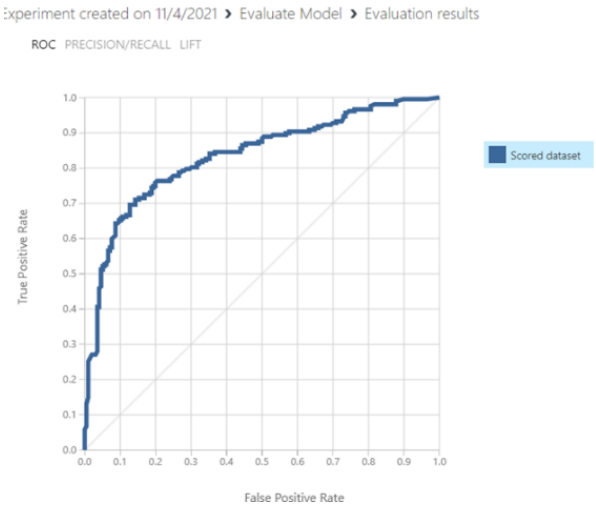

Figure 6 – Two-class Logistic Regression ROC curve

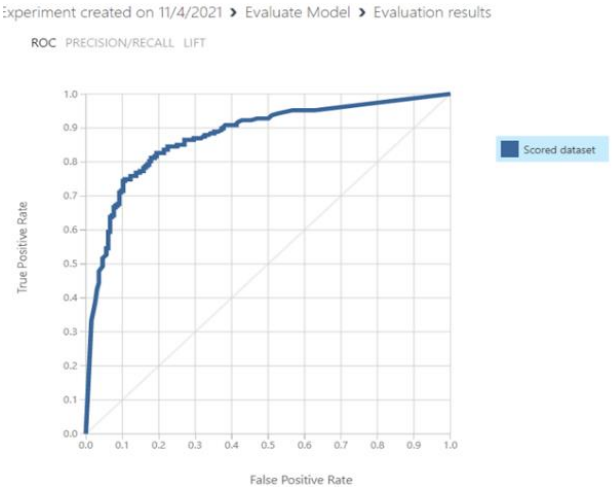

Figure 7 – Two-class Neural Network ROC curve

As previously stated, there are a total of 2017 Instances. The dataset is split 80-20 between train and test. As a result, we have 1613 occurrences of train data and 404 instances of test data. Our primary metrics are precision, recall, f1-score, and accuracy. We're building confusion matrixes and classification reports based on these numbers to see how effectively the model can distinguish between the two scenarios and how many false positives or false negatives it predicts during the process

Conclusion:

In the 21st century, smartphones have witnessed an exponential and almost 90% (7.1 billion) of the world population owns a smartphone. Almost each of these smartphones has a preinstalled music streaming app or a user can install a 3rd party application to stream music. The number of users on global streaming services is now at 490 million, which implies that the market is still open and can grow three folds. Music recommendation systems will play a major role in grabbing the customer base and in improving the overall experience.

Based on the data model we used these are the findings and our recommendation.

- Based on the accuracy given by the three models that we have used we concluded that the Two-class Neural Network has high accuracy when compared to the other 2 models and this would be helpful in predicting the songs appropriately. Considering this as a factor, we can enhance our customer base. This in turn would result in huge profits to the company.

- We want to introduce an application where the user would try this music prediction algorithm to get the songs of his choice by performing some analytics using the algorithms based on the data that we have.

Referencess

1. https://www.businesswire.com/news/home/20210803005884/en/24.7-Billion-Online-Music-Streaming-Market-by-Service-Revenue-Model-Platform-End-User-and-Content-Type---Global-Opportunity-Analysis-and-Industry-Forecast-2021-2027---ResearchAndMarkets.com
2. https://www.researchandmarkets.com/reports/5394229/online-music-streaming-market-by-service-revenue?utm_source=BW&utm_medium=PressRelease&utm_code=6gxg5p&utm_campaign=1571513+-+%2424.7+Billion+Online+Music+Streaming+Market+by+Service%2c+Revenue+Model%2c+Platform%2c+End+User%2c+and+Content+Type+-+Global+Opportunity+Analysis+and+Industry+Forecast%2c+2021-2027&utm_exec=chdo54prd
3. https://www.kaggle.com/geomack/spotifyclassification