Music Listening Behavior: Forecasting Track Skipping

Ghina Al Shdaifat GHA2009@NYU.EDU
Ansh Jhaveri AVJ2013@NYU.EDU
Kendall Brogle KLB9313@NYU.EDU

Milestone 2

1. Methodology

In this milestone, we conducted an in-depth investigation into the capability of machine learning algorithms, with a focus on neural networks, to predict song skipping behavior in a streaming music service context. We leveraged a dataset encompassing user-interaction logs and track features, which included both numerical and one-hot encoded categorical variables. To prepare for analysis, we merged the user log data with track feature data using a unique track identifier. We also extracted temporal features and transformed categorical attributes via one-hot encoding to ensure compatibility with machine learning models. Acknowledging the significance of temporal patterns in user behavior, we utilized Recurrent Neural Networks (RNNs) equipped with Long Short-Term Memory (LSTM) cells to model the probability distribution $P(y = 1|x = x_i)$. LSTMs were chosen for their proficiency in retaining information across extended sequences, which is vital in our context where previous user actions can influence the likelihood of skipping a track. In conjunction with deep learning techniques, we implemented traditional machine learning models to serve as baselines and points of comparison. These included: Random Forest Classifier with 100 trees and a fixed state, Gradient Boosting Classifier with 100 boosting stages, a learning rate of 0.1, and a maximum depth of 3 for the individual trees, and Logistic Regression to provide a baseline, with an interaction maximum of 1,000. These models were trained and tested on an 80-20 split dataset. We performed feature scaling using 'StandardScaler' to normalize the feature space and facilitate model convergence. For the neural networks, we employed cross-entropy loss as the objective function and used the Adam optimizer due to its adaptive learning rate capabilities. The performance of each model was meticulously quantified by tracking accuracy and plotting loss curves over successive training epochs. This allowed us to monitor the learning progression and adjust for overfitting or underfitting. Moreover, we extracted and analyzed feature importances from the trained models. This analysis was crucial for gaining insights into which features most strongly predict user skipping behavior, thereby enhancing our understanding of model behavior and informing potential service improvements. Through this methodological approach, we aimed to construct a predictive framework that not only performs with high accuracy but also provides interpretability and insights into user engagement within the music streaming service.

2. Results

Our investigation into song skipping behavior used various machine learning models, assessing their performance through statistical metrics. Accuracy, both in training and testing phases, represents the models' ability to correctly classify instances. Precision shows the model's proportion of true positive predictions in the total predicted positives. Recall assesses the model's ability to identify all actual positives. The F1-score is a single metric that balances both precision and recall. The AUC-ROC score is the measure of the ability of a classifier to identify between classes and is used as a summary of the model's performance. A feature importance score provides a measure of the influence each feature has on the prediction outcomes of a model.
The outcomes of the different models are as follows:

Targets	**Train Accuracy**	**Test Accuracy**	**Val Accuracy**	**Test Loss**
skip_1	84.91% – 88.44%	88.35%	∼88.69%	28.62%
skip_2	84.07% – 87.54%	87.61%	∼87.42%	26.94%
skip_3	95.75% – 98.15%	98.18%	∼98.32%	7.50%
not_skipped	95.89% – 98.82%	98.96%	∼98.77%	4.24%

Table 1: Performance metrics of LSTM RNN models for different targets predicting song skipping behavior.

Targets	**Accuracy**	**Precision (F/T)**	**Recall (F/T)**	**F1-Score (F/T)**
skip_1	87.80%	0.91/0.84	0.88/0.88	0.89/0.86
skip_2	87.33%	0.90/0.85	0.83/0.91	0.86/0.88
skip_3	98.10%	0.99/0.98	0.96/0.99	0.97/0.99

Table 2: Performance metrics of the Random Forest model on predicting different skip behaviors, including both the positive (T) and negative (F) classes.

Targets	**Accuracy**	**Precision (F/T)**	**Recall (F/T)**	**F1-Score (F/T)**
skip_1	88.12%	0.92/0.84	0.87/0.89	0.89/0.86
skip_2	87.53%	0.90/0.86	0.84/0.91	0.87/0.88
skip_3	98.21%	0.99/0.98	0.95/1.00	0.97/0.99

Table 3: Performance metrics of the Gradient Boosting model on predicting different skip behaviors, including both the positive (T) and negative (F) classes.

Targets	AUC-ROC Score	Accuracy
skip_1	93.88%	87.80%
skip_2	94.46%	87.33%
skip_3	98.88%	98.10%

Table 4: Performance metrics of the Logistic Regression model on predicting different skip behaviors.

Feature	Random Forest	Gradient Boosting	Logistic Regression
End Trackdone	0.186/0.286/0.474	0.595/0.828/0.975	-2.09/-2.08/-2.58
Start Trackdone	0.115/0.084/0.071	0.214/0.086/NA	-0.86/-0.56/NA
End Fwdbtn	0.106/0.114/0.186	0.074/0.113/0.013	1.16/1.12/1.45

Table 5: Comparison of top feature importances across Random Forest, Gradient Boosting, and Logistic Regression models for Skip 1, 2, and 3 scenarios, recorded in this order.

3. Analysis

Evaluating various machine learning models shows a comprehensive learning of user interactions with a music streaming service, particularly in relation to song skipping behavior. For early skips (‘skip 1‘ and ‘skip 2‘), both the RNN and Random Forest models achieved high accuracy during training and testing, suggesting a strong capability in recognizing factors influencing a user’s decision to skip at the beginning of a track. However, it is worth noting that there is a considerable loss for these early skips when compared to late skips (‘skip 3‘) and tracks not skipped at all (‘not skipped‘). This disparity could result from more complexity in user behavior patterns associated with early skips. In contrast, the performance on ‘skip 3‘ is high across all models, with the RNN model showing particularly high validation accuracy and minimal test loss. These results suggest that user behaviors leading to late skips are more predictable, potentially due to more discernible patterns because the user has listened to more of the track before deciding to skip. It could also be an indicator of overfitting or data imbalance. The ‘not skipped‘ behavior obtained near-perfect accuracy in the RNN model, indicating a robust understanding of the characteristics of tracks that retain user engagement until the end of the song. This is a critical insight, as it may reflect user satisfaction or a positive listening experience. The Logistic Regression model, regardless of its simplicity, had a high AUC-ROC score across all skip behaviors, particularly for ‘skip 3‘. Across all models, the most influential features for predicting skips relate to user actions at the end of track playback. The features indicating the reasons for track playback ending (‘End Trackdone‘) and starting (‘Start Trackdone‘) are significant predictors in the Random Forest and Gradient Boosting models, especially for early skips. The logistic regression model highlights the importance of these features, particularly for ‘skip 3‘.

4. Plan for Additional Analysis

Although we have implemented a fairly straightforward/simple LSTM RNN model and our values are very promising, we could improve our research methodology by conducting further research to ensure accurate reflections of user skip behavior.
Since there are concerns about overfitting, especially for variables 'skip 3' and 'not skipped', we could consider simplifying the model or introducing regularization techniques. Additionally, the values used in our project is a smaller version of the full dataset due to storage and memory constraints. It could be beneficial to cross-validate across different data splits to ensure the models' robustness and determine a more accurate understanding of the models' overall performance on unseen data. One additional element we aim to implement by the Milestone 3 deadline, which we believe could be interesting, is to use a time-series LSTM model where we can observe the progress of time (in years) to predict the prevalence of the four most popular and common features identified during classification in the next 5 years. This will allow us to understand whether there is a specific feature in a song that is highly correlated with skip behavior to further our understanding of not only the progression of music over time, but also certain trends and patterns of skip behavior throughout the modernization of songs.
An additional aspect we plan to add is word embedding to analyze lyrics. To accomplish this, we will use Natural Language Processing (NLP) techniques to convert lyrics into numerical data that our models can process. This will allow us to include lyrical content as features in our models and examine whether the themes, sentiments, or specific word choices in lyrics influence the likelihood of a song being skipped.
By embedding the lyrics, we hope to find patterns within the text that may correlate with skip behavior. To analyze this, we will integrate the lyrical embeddings into our LSTM RNN model, allowing it to consider both the acoustic features of the song and the lyrical content.

5. Work Plan

Leading up to this milestone our team has worked to communicate and divide work. Overall, Ansh took the lead on implementing the machine learning techniques, including the random forest, gradient boosting and logistic regression models and their feature importances. Ghina took charge of the RNN modeling to use sequential patterns in her predictions. Kendall has been working on adding the lyrical encoding element that we plan to add to the next milestone as well as the written work for the milestone. We met frequently, working together on most aspects of the work, making it a largely collaborative result.
In the next week Kendall will finalize the lyrical embedding work with help from Ansh. Ghina will work on refining her RNN models and implementing the time series analysis. We will then meet to analyze our final results and begin the writing process for our final milestone. After this, Ghina will compile the individual videos we will each take of our individual findings. We will all make slides pertaining to our individual contributions. Kendall will be in charge of uploading the GitHub repository. We will continue to meet a few times a week to receive feedback and help from each other.

Appendix

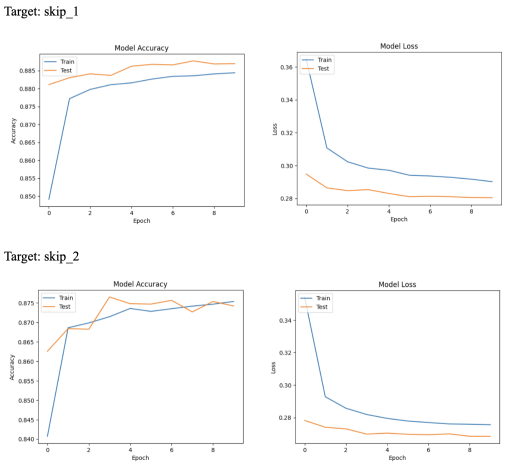

Figure 1. Model Accuracy and loss for skip 1 and skip 2

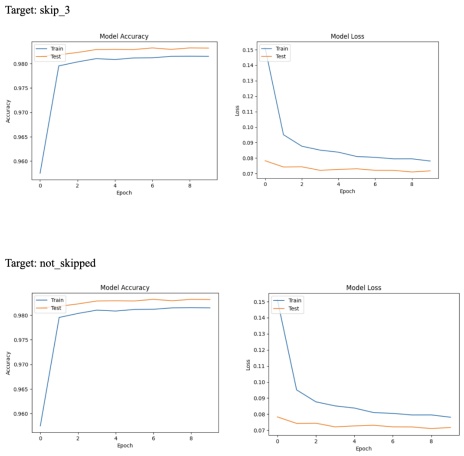

Figure 2. Model Accuracy and loss for skip 3 and no skip

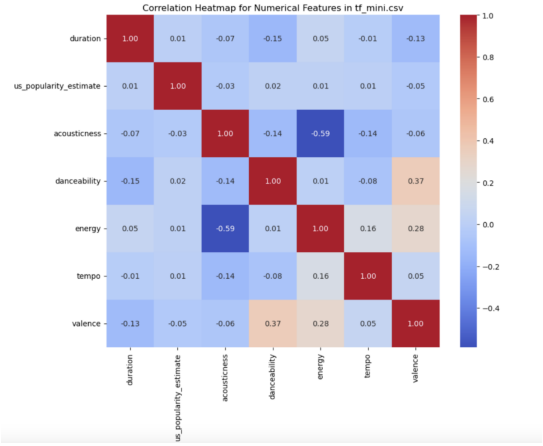

Figure 3. Heatmap for numerical features

References

[1] Sainath Adapa. 2019. Sequential modeling of Sessions using Recurrent Neural Networks for Skip Prediction. arXiv preprint arXiv: 1904. 10273 (2019)
[2]Elbir, A. and Aydin, N. (2020), Music genre classification and music recommendation by using deep learning. Electron. Lett., 56: 627-629. https://doi.org/10.1049/el.2019.4202.