Music Listening Behavior: Forecasting Track Skipping

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

Milestone 3

1. Methodology:

In this milestone, we are conducting an in-depth analysis of neural network models, specifically by implementing hyperparameter optimization. The focus of this milestone was on two key variables: skip 1, which indicates that a track was skipped almost as soon as it has been played, and skip 3, which indicates that a user has skipped near the end of the track but not to completion, suggesting a more engaged interaction. These variables serve as representations for user satisfaction and engagement, with the assumption that skip 3 and not skipped (where a track is played to completion) may act as "ground truth" indicators of user contentment with the track selection, while skip 1 and skip 2 indicate skips happening early on through the track. While skip 2 suggests that the track has only been played for the first 30 seconds, we have decided to use skip 1 and skip 3 for a more extreme comparison to identify the effect of user skip behavior at the start and nearly the end of a track.

In milestone 2, the implementation of a simple Long Short-Term Memory (LSTM) models were created for each individual "skip" variable. Initially, the LSTM models were trained on a dataset partitioned into 70% training and 30% evaluation subsets, using a standard batch size of 64 and a learning rate of 0.001, with the aim of capturing temporal sequences in user interaction data. Overall, the initial models indicated a disparity in predictive performance, with skip 1 and skip 2 showing consistent accuracy across training, validation and testing datasets, while skip 3 and not skipped demonstrated overall higher accuracies.

To unravel the complexities of skip 1 and skip 3, hyperparameter tuning was introduced into the analysis using Keras Tuner's Hyperband algorithm. This approach employs resource allocation strategies and early-stopping techniques. The Hyperband was configured to iterate over a maximum of 10 epochs for each configuration, reducing the number of epochs by a factor of 3 at each bracket and exploring hyperparameters such as LSTM units (ranging from 32 to 256) and dropout rates (tested from 0.0 to 0.5 in increments of 0.1). By doing so, we were able to fine-tune the model's learning capacity and address potential overfitting.

Following the tuning phase, we decided to refine the LSTM model architectures. Dropout layers (dropout rate of 0.3) were incorporated to prevent the models from becoming overly reliant on any singular patter within the training data, and and the LSTM units were optimized to 128 to balance model complexity and training efficiency. Additionally, batch normalization layers were introduced to regulate the internal covariate shift, leading to a more stabilized learning process by normalizing the activations. Dropout layers (dropout rate of 0.3) were additionally incorporated to prevent the models from becoming overly reliant on any singular patter within the training data. Training the refined models on

the full dataset, an early stopping mechanism with a patience parameter of 10 epochs was employed. Alongside, the ModelCheckpoint callback was used to save the best model based on the lowest validation loss observed. The evaluation of the models' predictive power was quantified through the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) metric, offering a clear indication of their true positive rate against the false positive rate.

Finally, the models' decision-making processes were investigated using SHAP values, which provided a mean(|SHAP value|) ranging from 0.01 to 0.14, offering insights into the impact of features on the prediction outcomes. Therefore, the general methodology of this milestone, not only enhances the performance of the predictive models but also provides a more specific understanding of the factors that influence user behavior in the music streaming field.

2. Results:

The provided results for the skip1 and skip 3 LSTM models reveal distinct predictive performances. For skip 1, the model achieved a peak accuracy of about 88% and demonstrated early stopping around epoch 17 (Figure 1), indicated by the dashed line in the accuracy graph. The most influential feature for the skip 1 model, as indicated by the SHAP value plot, is 'hist user behavior reason end endplay' (Figure 2), remains consistent with the skip 3 model (Figure 6), implying its strong predictive power across different types of skip behavior. This feature's prominence in both models denotes the user's action of allowing a track to end normally, implying a complete listening session without skipping. Its high impact suggests that when users typically listen to tracks in their entirety, it is a strong indicator of their future behavior regarding longer engagement with a track before deciding to skip. Both models exhibit high Receiver Operating Characteristic (ROC) curve areas, with skip 1 at 0.94 (Figure 3) and skip 3 (figure 5) at 0.99, indicating excellent discrimination between the positive and negative classes for both models. However, the higher AUC for skip 3 reflects an even stronger predictive capability for this model, likely due to the more definite user engagement it represents.

3. Analysis:

The skip 1 model, with an accuracy of around 88%, suggests a well-trained network that generalizes well to unseen data. The most influential feature, "hist user behavior reason end endplay", indicates that the model has learned that users who let the previous track play to the end are less likely to skip the next track within the first few seconds. This suggests a level of user satisfaction with the platform's track sequencing or reflects full-track listening before initiating a skip. This is further supported by skip 1's high level of predictive accuracy demonstrated by the ROC/AUC curve, allowing for confident differentiation between instances of skipping and not skipping within the first moments of a track.

The skip 3 model demonstrates exceptional accuracy, reaching up to 98.2%. The model's quick convergence and low loss indicate that skip 3 behavior can be predicted with high confidence, potentially due to more pronounced patterns in user engagement for this behavior. The dominant feature, "hist user behavior reason end endplay" (further supported in our

ML analysis), remains significant, underscoring that the model has effectively captured the tendency of users to skip a track after more extended engagement based on their previous behavior of listening to tracks in full. An impressive ROC/AUC curve for skip 3 suggests the model has a superior capability to distinguish between the nuanced behavior of skipping after longer engagement, compared to the skip 1 model.

Overall, the comparison between skip 1 and skip 3 models reveals that user behavior regarding the completion of previous tracks is a strong predictor of subsequent skipping behavior, whether immediately or after some engagement.

4. Conclusions:

Overall, the models were adept at discerning user engagement patterns, achieving high ROC/AUC scores. A pivotal finding was the significance of "hist user behavior reason end endplay" across both models, indicating that the completion of the previous track is a strong predictor of future skipping actions, which is further supported by our ML analysis. Despite these strengths, a key weakness emerged: the models' heavy reliance on this single feature raises concerns about their adaptability to diverse and evolving user behaviors. Moreover, the high accuracy scores, particularly for the skip 3 model, suggest a need to investigate for potential overfitting, although regularization techniques were in place.

The project, while successful, has limitations, including potential dataset biases and a lack of deeper feature analysis that could unveil more information on user interactions. Future work could involve integrating more contextual variables, exploring robustness across different user groups, and considering multi-objective optimizations for hyperparameter tuning.

Interestingly, despite incorporating complexities into the LSTM models to refine their predictive power, we observed only marginal improvements in accuracy compared to simpler LSTM models implemented in Milestone 2. This raises a practical consideration about computational costs and time efficiency, without significantly compromising predictive performance, making them a compelling choice for real-world application in a fast-paced music streaming environment.

5. Workflow:

During the first milestone, Ansh was responsible for conducting the EDA, while Kendall and Ghina were responsible for conducting research, exploring potential methods and techniques. By milestone 2, Ansh was responsible for conducting further ML analysis, while Ghina implemented DL techniques. Kendall's role additionally involved conducting a word embedding system using NLP to identify words in certain tracks and essentially discover whether certain skip behaviors are correlated to specific words. Unfortunately, we were unable to get the song ids and faced difficulty in gaining access to the Spotify API in time. By milestone 3, Ghina conducted further analysis on hyperparameter tuning techniques on the LSTM models, while Ansh and Kendall aided in writing the final milestone. Moving forward, all three of us are working together on the presentation.

Appendix

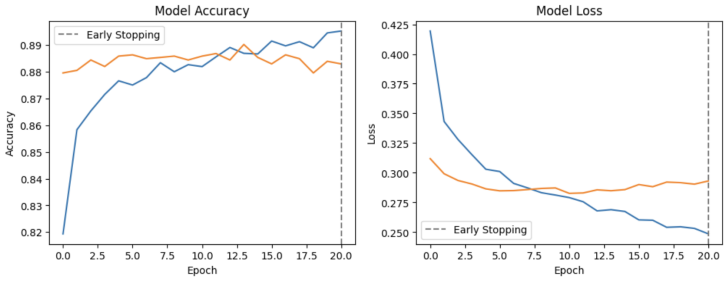

Figure 1: Model Loss and Accuracy Graphs for Skip 1

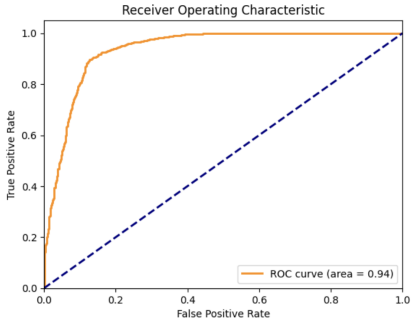

Figure 2: ROC for Skip 1

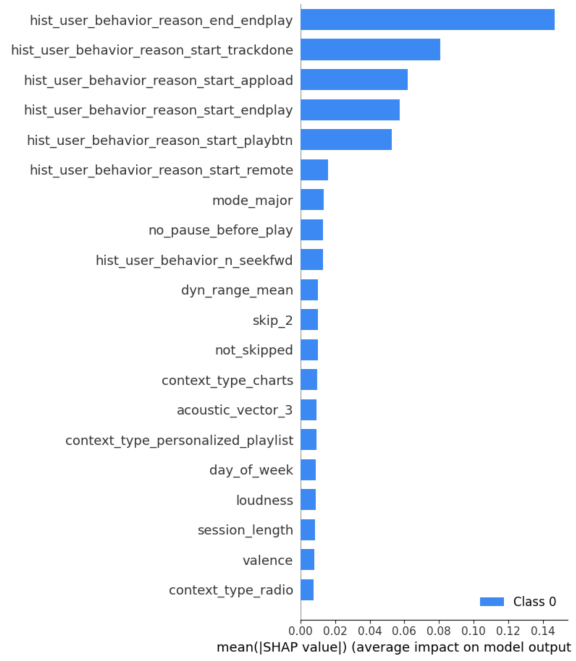

Figure 3: Influential Features for Skip 1

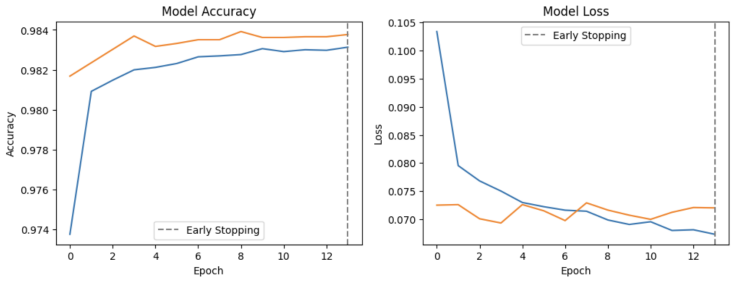

Figure 4: Model Loss and Accuracy Graphs for Skip 1

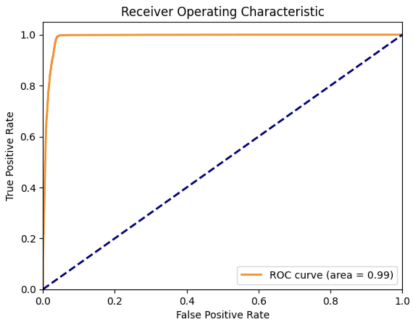

Figure 5: ROC for Skip 1

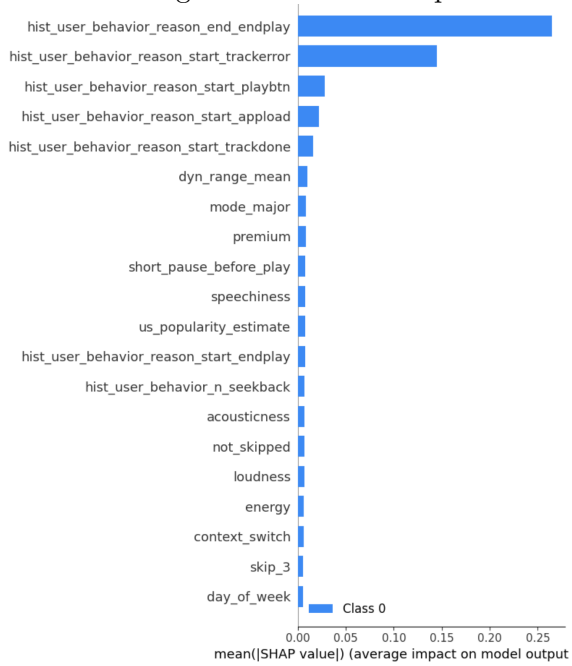

Figure 6: Influential Features for Skip 1