System Analysis Report

Daniel David Cuellar, Henry Jhonmarcos Osorio Sánchez,
Juan Andrés Jimenez Palomino, Miguel Alejandro Chavez Porras

Universidad Distrital Francisco Jose de Caldas

Abstract. In this report we are going to analyze, do an overview of all the system, see its interactions between the parts that compose the system and do some research to define the complexity, sensitivity, chaos and randomness in this system.

Keywords: Music · System Analysis · Prediction model

1 Introduction

The KKBOX Music Recommendation Challenge (WSDM 2018) addresses a real-world problem faced by music streaming platforms: predicting whether a user will listen to the same song again after the first observable listening event. Accurate prediction of repeated listening behavior is essential for improving recommendation quality, enhancing user engagement, and increasing customer retention. The competition provides a rich dataset with millions of user-song interactions, including user demographics, song metadata, and contextual information of how the listening event was triggered. Unlike simplified toy datasets, this competition simulates the complexities of real streaming services, where unpredictable user behavior, noisy metadata, and dynamic temporal patterns introduce significant modeling challenges.

2 Objetives

Predict the probability that a user will listen to the same song again within one month of the first observable listening event. The target variable is binary: 1 for repeat listening, 0 otherwise.

3 Analysis

The system consists of four key elements: users, songs, listening events, and the supporting dimensions of time and context. Listening events serve as the central point of interaction, connecting users with songs.

- Users contribute demographic and registration attributes that can shape their listening patterns.

- Songs provide metadata such as artist, genre, length, and language, which influence user preferences.
- Listening events record how and when a user interacted with a song, enriched with contextual variables such as application tab or entry type.
- Time determines whether a repeat listen occurs within the one-month window required by the target variable.
- Context provides information about the environment of the interaction (search, playlist, library), which can affect the likelihood of repetition.

The relationships between these elements can be described as follows:

- Each user generates multiple events.
- Each song is associated with many events
- Every event is tied to a time dimension and occurs under a context.

This systemic view emphasizes that the listening event is the dynamic hub of the system, while users and songs act as static entities enriched with attributes.

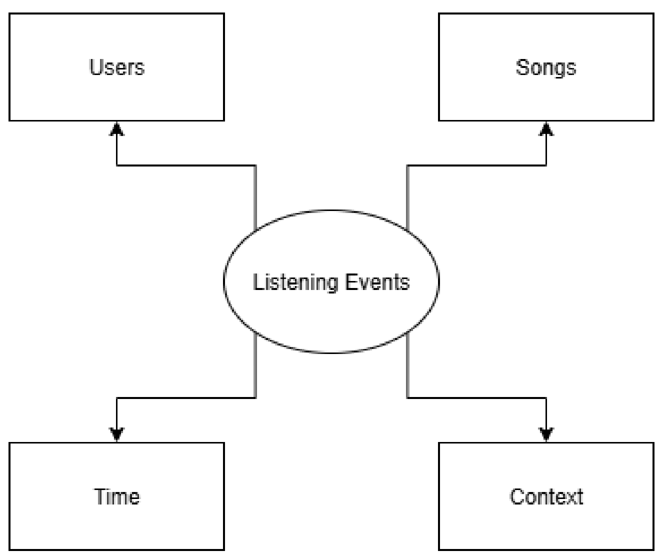

Fig. 1. System elements and relationships.

4 Complexity and Sensitivity

The system exhibits multiple sources of complexity, primarily due to the high dimensionality of the data, the presence of outliers, and the nonlinear relationships among variables. A key part of the analysis is to understand how sensitive the system is to changes in its components:

– Feature Sensitivity: The model's predictive performance can vary significantly depending on which features are included. For instance, removing genre_ids or misinterpreting isrc codes reduces accuracy, while contextual features such as source_type often improve predictions.
– Cold-Start Sensitivity: When new users or songs appear with no historical data, predictions become highly unstable. Small changes in demographic attributes or song metadata can disproportionately affect outcomes.
– Noise in Input Variables: Extreme values in the bd (age) column, if not filtered, can mislead the model and produce biased probability estimates.
– Temporal Sensitivity: Since train and test datasets are split by time, shifts in user preferences over periods can degrade model generalization.

Methods to analyze sensitivity include:

– Feature importance ranking (e.g., SHAP values) to quantify the impact of each variable.
– Ablation studies where groups of features are removed to observe performance degradation.
– Perturbation experiments introducing small random changes in inputs to evaluate stability.

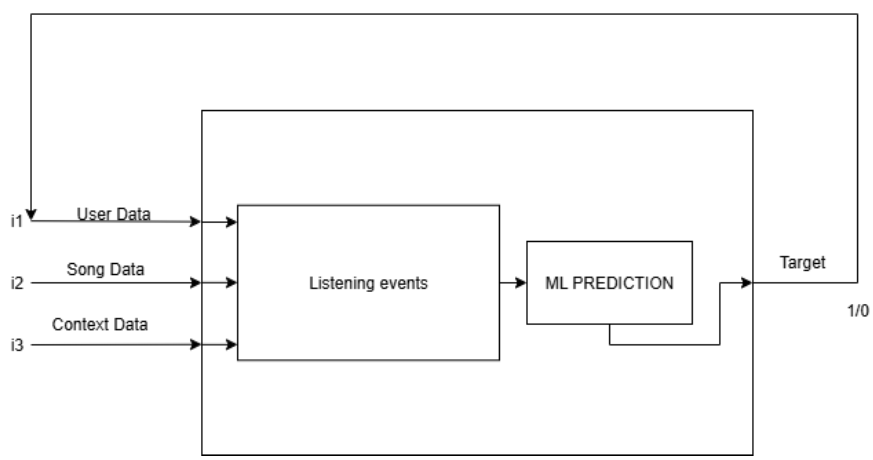

Fig. 2. System Architecture.

Available Data

1. train.csv: msno, song_id, source_system_tab, source_screen_name, source_type, target.

2. test.csv: id, msno, song_id, source_system_tab, source_screen_name, source_type.
3. songs.csv: song_id, song_length, genre_ids, artist_name, composer, lyricist, language.
4. members.csv: msno, city, bd (age, with outliers), gender, registered_via, registration_init_time, expiration_date.
5. song_extra_info.csv: song_id, song_name, isrc.

Evaluation Metric
Area Under the Receiver Operating Characteristic Curve (AUC) between predicted probabilities and observed targets.
Constraints and Notes

1. Train/test split is strictly time-based.
2. ISRC codes may contain noise and inconsistencies.
3. Several columns (e.g., age) include extreme outliers.

5 Elements and relationships

System Elements

1. Users: demographic and registration attributes (city, age, gender, registration time).
2. Songs: metadata such as length, genre, artist, composer, lyricist, and language.
3. Events: contextual information of listening (system tab, screen, entry type).
4. Target: binary label indicating whether a re-listen occurred within one month.

Elements relationship

1. User → Events: each user generates multiple listening events.
2. Song → Events: each song can be associated with many different users.

System boundaries

1. Temporal division between training and test datasets.
2. Limited to observed interactions within the KKBOX service.

6 Chaos and Randomness

1. Unpredictable User Behavior: users do not follow stable patterns; preferences may shift abruptly without clear correlation.
2. Cold-Start Problem: recommendations for new users or songs lack historical context, increasing uncertainty and sensitivity to minor variations.
3. Feedback Loops: recommendations influence future listening, reinforcing certain genres and distorting the observed dataset distribution.
4. Noisy and Inconsistent Data: age values may be unrealistic; ISRC codes are sometimes duplicated or misassigned, producing irregularities.

5. Nonlinear Interactions: outcomes depend on complex feature interactions (e.g., genre $\times$ age $\times$ device $\times$ source type), where small input changes can lead to large prediction shifts.
6. Unmodeled External Events: new releases, viral phenomena, or concerts may drastically affect listening behavior but are absent from the dataset.
7. Dynamic Temporal Distribution: user preferences evolve over time (seasonality, trends, holidays), making models trained on one period quickly outdated.

7 Conclusion

The analysis of the KKBOX system highlights both its strengths and challenges. The dataset is rich and diverse, enabling exploration of user preferences across demographics, song attributes, and contextual features. However, the system is inherently complex: user behavior is unpredictable, the data includes noise and inconsistencies, and the temporal dimension introduces instability in predictions. From a systems engineering perspective, the challenge demonstrates how elements such as users, songs, events, time, and context interact dynamically to produce outcomes. Sensitivity analysis confirms that model performance is highly dependent on careful feature selection and preprocessing. Finally, the presence of chaotic and nonlinear aspects, such as feedback loops and cold-start conditions, emphasizes the need for robust modeling techniques. In conclusion, while the KKBOX dataset provides an excellent opportunity for experimentation with advanced recommendation algorithms, successful solutions must account for sensitivity, complexity, and chaotic dynamics that characterize real-world systems.

References

1. https://www.kaggle.com/competitions/kkbox-music-recommendation-challenge/overview
2. https://github.com/EngAndres/ud-public/tree/main/courses/systems-analysis
3. Systems Analysis and Design, by Alan Dennis, Barbara Haley Wixom, and Roberta M. Roth