Evaluation of SHAP for Songs dataset

Miguel and Gian Paolo

USC Master in Artificial Intelligence

1 Introduction

The growing use of machine learning models in sensitive domains has increased the demand for Explainable Artificial Intelligence (XAI). While many explanation methods have been proposed, evaluating their effectiveness remains a key challenge. An explanation is only useful if it is understandable and meaningful to its intended users.

The goal of this practical session is to design and conduct a user study to evaluate the quality of explanations provided by a selected XAI system. We implement a human-grounded evaluation involving experts in XAI, using a structured questionnaire to assess clarity, usefulness, and trust. The results are analyzed statistically to validate the proposed research hypotheses.

2 XAI System Under Evaluation

In this study, we evaluate the SHAP (SHapley Additive exPlanations) method applied to a classification model trained on the SONGS dataset, which predicts whether a song will be classified as "Like" or "Dislike" based on audio features.

2.1 Model Description

The underlying predictive model is a supervised classifier trained to predict the likelihood of the song based on some music attributes. SHAP is used as a post-hoc explanation method to quantify the contribution of each feature to the model's predictions, both at the global and local levels.

2.2 Explanation Method

SHAP explanations are presented through visual plots, such as waterfall plots. These visualizations show the magnitude and direction of each feature's contribution to the final prediction.

3 Evaluation Design

3.1 Type of Evaluation

This study follows an **intrinsic evaluation**, as the goal is to assess the quality of the explanations themselves rather than their impact on performance in an external task. We focus on properties such as clarity, interpretability, and perceived usefulness of the explanations generated by SHAP.

In addition, the evaluation is primarily **quantitative**, since participants will respond to structured questionnaire items using Likert scales. This allows us to perform statistical analysis to validate the proposed research hypotheses.

3.2 Assessment Type

We adopt an **application-grounded assessment**, as the evaluation is conducted with experts in XAI who assess explanations in the context of a real predictive problem (song preference classification). Since the target population consists of knowledgeable participants, this setting allows for a realistic evaluation of explanation quality in the context of music preference prediction.

4 Target Population

The target population of this study consists of a panel of experts in Explainable Artificial Intelligence (XAI). Specifically, participants are classmates enrolled in the MIA program who have prior knowledge of machine learning models and explanation methods, including SHAP, Decision Trees, and Fuzzy Rule-Based Systems.

- **Expertise:** Participants have academic training in XAI concepts and practical experience working with interpretability techniques during previous practical sessions.
- **Background:** The majority have a technical background in Artificial Intelligence, Data Science, or related fields, ensuring familiarity with classification models and explanation tools.
- **Recruitment process:** Once the questionnaire is implemented in Microsoft Forms, it will be shared with colleagues from other MIA teams through the course communication channels.

5 Samples and Explanations Evaluated

For the evaluation study, we selected four specific data instances from the SONGS dataset: instances 228, 327, 150 and 200. These songs were chosen to represent contrasting model predictions ("Like" and "Dislike"), allowing participants to assess explanations in different decision scenarios.

All available audio features used by the model were displayed to participants, including: acousticness, danceability, duration_ms, energy, instrumentalness, key, liveness, loudness, mode, speechiness, tempo, time_signature, and valence. These 13 attributes correspond to the musical characteristics provided in the dataset and constitute the feature space used for training the classifier.

Identifier variables such as song name, artist, and any instance ID were not used as input features in the model and were therefore not considered in the explanations presented to participants. Providing the complete set of relevant audio attributes ensures that the explanations are interpreted within the actual feature space used by the Decision Tree classifier, avoiding potential bias from non-predictive metadata.

The explanations were presented using SHAP **Waterfall plot** (local explanation) to illustrate how individual feature contributions combine to produce the final prediction.

These different explanation formats allow us to evaluate both global understanding of the model behavior and local interpretability for individual predictions.

6 Research Questions and Hypotheses

6.1 Research Questions

The study aims to answer the following research questions:

- **RQ1:** Do SHAP waterfall plots enable participants to correctly interpret the model's prediction (predicted class and influential features)?
- **RQ2:** How clearly are SHAP waterfall plots perceived by participants?
- **RQ3:** Do SHAP explanations increase participants' trust in the model's predictions?

6.2 Hypotheses

Based on these research questions, we propose the following hypotheses:

- **H1:** Participants will correctly identify the predicted class and most influential features at a rate significantly above chance level.
- **H2:** SHAP waterfall plots will receive clarity ratings significantly above the neutral midpoint of the Likert scale.
- **H3:** Participants will report higher levels of trust in the model's predictions when SHAP explanations are provided.

7 Evaluation Plan

The evaluation will be conducted through a structured online questionnaire implemented in Microsoft Forms (Quiz mode). The questionnaire is organized into sections. Participants will first complete a short comprehension test based on one SHAP waterfall plot, followed by three additional task-based sections where they interpret different plots corresponding to individual song instances.

For each song, participants will be presented with:

- The SHAP waterfall plot explaining the Decision Tree prediction (local explanation).
- Objective questions assessing their understanding of:
  - The most likely predicted class ("Like" or "Dislike").
  - The most influential features.
  - The direction of influence (positive or negative contribution).
- Subjective questions evaluating the perceived clarity and usefulness of the explanation.

Tasks:

- Interpret each SHAP waterfall plot.
- Identify the most influential features and their direction of contribution.
- Infer the most likely predicted class based on the explanation.
- Provide a global assessment of the interpretability of the explanations.

Metrics:

- *Objective comprehension metrics:* Accuracy in identifying the predicted class, most influential features, and direction of contribution (automatically scored in Quiz mode).
- *Subjective perception metrics:* Likert-scale (1–5) ratings measuring clarity, usefulness, ease of understanding, and trust in the model.
- *Qualitative feedback:* Short open-ended responses regarding strengths and potential improvements of the explanations.

Data Collection Process: The explanations are generated using a classification model trained on the SONGS dataset. Feature values for 2017 songs are contained in `SONGS.arff`, while song names and IDs are stored in `spotifyData.csv` to identify the instances presented to participants. The generated plots are embedded into the questionnaire, and responses are collected anonymously through Microsoft Forms for subsequent statistical analysis.

8 Data Management Plan

This project involves two types of data: (1) the SONGS dataset used to generate model explanations, and (2) the responses collected from participants during the evaluation study.

- **Data Collection:** The SONGS dataset (SONGS.arff and spotifyData.csv) is publicly available from a Kaggle competition and contains audio features and song identifiers. Participant responses are collected through Microsoft Forms in the form of questionnaire answers.
- **Data Storage:** The dataset files and generated explanations are stored locally on secure university computers. Questionnaire responses are stored securely within the Microsoft Forms platform for statistical analysis.
- **Anonymization:** No personal identifiers (names, emails, student IDs) are collected from participants. Responses are analyzed in aggregated form to ensure anonymity. The SONGS dataset does not contain sensitive personal data.
- **Data Retention:** Collected data will be retained only for the duration necessary to complete the assignment and grading process. After project completion, all participant response files will be deleted in accordance with good data management practices.

9 Ethics Committee Approval Process

Although a formal submission to the Ethics Committee was not required for this academic assignment, we reviewed the official documentation provided by the Research Ethics Committee of the University of Santiago de Compostela (Comité de Ética en Investigación da USC) to familiarize ourselves with the approval process and required information.

According to the institutional guidelines, any research project involving human participants, biological samples, or personal data must submit a formal application entitled *"Solicitude de informe ao Comité de Ética en Investigación da USC"*. The application must be sent to the committee's official email address and includes the following main components:

- **General Information:** Identification of the principal investigator, institutional affiliation, project title, duration, and a summary of the research (maximum 250 words).
- **Research Team Qualifications:** Academic background, institutional relationship, specific tasks, and relevant experience of each team member involved in data collection or participant interaction.
- **Scientific and Methodological Aspects:** Description of the research objectives, justification of the study, methodological design, instruments used (e.g., surveys), and planned data analysis procedures.
- **Use of Human Participants and Personal Data:** Justification for involving human subjects, description of recruitment procedures, type of intervention (e.g., surveys or questionnaires), data protection measures, and compliance with GDPR regulations.
- **Informed Consent Documentation:** Clear explanation of voluntary participation, anonymity or confidentiality guarantees, purpose of data collection, and participants' rights (including withdrawal).
- **Additional Documentation:** Recruitment materials, consent forms, and any supporting scientific references relevant to the project.

In the context of this study, participants only completed an anonymous online questionnaire implemented in Microsoft Forms. No biological samples were collected, and no sensitive personal data were processed. The only data gathered consisted of questionnaire responses and optional demographic information (e.g., age range, academic background), which were collected anonymously and used exclusively for academic research purposes.

The study involves minimal risk, as participants are asked solely to interpret SHAP waterfall plots and provide their opinions regarding clarity and usefulness. Participation is voluntary, no compensation is offered, and no identifying information (such as names or email addresses) is automatically collected.

Although formal ethics approval was not requested for this coursework activity, the questionnaire was designed in accordance with the principles outlined in the Ethics Committee documentation, particularly with respect to informed consent, data minimization, anonymity, and compliance with the General Data Protection Regulation (GDPR).

10 Questionnaire Implementation

The questionnaire was implemented using **Microsoft Forms** in Quiz mode in order to enable automatic scoring of objective comprehension questions. The form is organized into multiple sections to ensure a structured and progressive evaluation process.

The first section consists of a short *comprehension test* based on a single SHAP waterfall plot, designed to verify that participants understand how to interpret local explanations. The following three sections each present a different SHAP *waterfall plot* corresponding to an individual song instance. For each plot, participants are asked to:

- Identify the most influential features contributing to the prediction.
- Infer the most likely predicted class based on the explanation.
- Determine whether specific features push the prediction toward "Like" or "Dislike".

In addition to objective comprehension questions, subjective evaluation items are included using 5-point Likert scales to measure:

- Clarity of the explanation.
- Perceived usefulness.
- Ease of understanding.
- Trust in the model's prediction.

The questionnaire concludes with a global assessment section and a short demographic survey (gender, age range, university, English proficiency) to support subsequent statistical analysis.

10.1 Platform Limitations

While Microsoft Forms is user-friendly, it has some limitations that affect the implementation of best practices for XAI evaluation:

- **Limited customization of visualizations:** Plots must be embedded as static images, so interactive features (like zooming or highlighting features dynamically) cannot be implemented.
- **No adaptive questioning:** Conditional logic is limited, preventing dynamically tailoring questions based on participant responses.
- **Restricted data export:** While responses can be exported to Excel, more advanced analytics (like automated scoring of open-ended feedback) requires manual processing.
- **No integration of external code:** Direct interaction with Python or other model outputs inside the form is not possible; all explanations must be pre-generated.

11 Pretest

A pretest of the questionnaire was conducted with our team members (Miguel and Gian Paolo) to ensure clarity, usability, and completeness. Each of us completed the questionnaire independently, simulating the experience of an external participant.

During the pretest, we identified the following issues and implemented improvements:

- **Clarification of instructions:** Some questions about feature importance were ambiguous, so we added explanatory text and examples.
- **Plot visibility:** The embedded SHAP plots were resized and labeled more clearly to improve readability.
- **Ranking questions:** Initially, the ranking options were confusing; we simplified the wording and ordering.
- **Likert scale consistency:** Some scales had inconsistent ranges; all Likert questions were standardized to a 1–5 scale.

After these adjustments, the questionnaire was deemed ready for deployment to the target population.

12 Evaluation Study

The evaluation study was conducted following the finalized questionnaire and evaluation plan. Details include:

- **Number of participants:** Specify the total number of participants who completed the survey.
- **Duration:** Indicate the time span during which the study was conducted.
- **Deviations from the plan:** Report any changes or unexpected issues encountered during the study execution.

Participants were asked to complete the tasks using the provided SHAP waterfall plots, and their responses were collected for statistical analysis.

13 Statistical Analysis and Results

13.1 Analysis Methods

Describe the statistical methods and tests used to analyze the collected data, for example:

- Descriptive statistics (mean, median, standard deviation)
- Likert scale analysis (e.g., average ratings, distributions)
- Comparative tests (e.g., paired t-tests, Wilcoxon signed-rank test) for differences between explanation formats
- Correlation or regression analyses, if relevant

13.2 Results

Present the results using tables and figures. Include:

- Average ratings for each explanation format
- Rankings of clarity, usefulness, and trust
- Observed patterns or trends across participants

13.3 Hypothesis Validation

State whether the research hypotheses are accepted or rejected based on the analysis results. Provide justification, for example:

- If the hypothesis predicted that "high values of danceability, energy, and mode imply a song is liked," summarize the findings and indicate whether the participant responses support this.
- Explain any unexpected results or limitations affecting hypothesis validation.

14 Discussion

Discuss:

- Interpretation of results
- Limitations
- Threats to validity

15 Conclusions

Summarize findings and suggest future work.