Evaluation Report

Absenteeism Prediction System

Krishna Kumar Bais (241110038)
Rohan(241110057)

1 Summary

This document reports a systematic evaluation of the final prototype user interface and explainability components for the Absenteeism Prediction System. The evaluation covers (a) an assessment of the interface and explainability features, (b) the evaluation methodology used, (c) findings, (d) recommended improvements, (e) reflections on the method itself, and (f) evidence. The evaluation is based entirely on structured think-aloud sessions and usability testing with observations.

2 (a) Evaluation of the interface and explainability features

2.1 Scope of evaluation

We evaluated the following components of the prototype:

- Input form layout and information architecture.
- Prediction presentation (numeric result, confidence).
- Explainability modules: SHAP (local + global), LIME (local), counterfactual explanations, model transparency panel (performance, fairness).

2.2 Key UI/UX observations

The interface uses a clean layout with clear grouping. Several features are aligned with good UI/UX practices, but the evaluation identified issues related to labeling clarity, discoverability of explanations, and perceived layout efficiency.

2.3 Explainability features (technical summary)

The prototype includes:

- SHAP for global and local explanation.
- LIME as a local surrogate explanation method.
- Counterfactual explanations for minimal actionable changes.

Design choices:

- Top-5 feature contribution list.
- Cached global explanations to reduce latency.

3 (b) Evaluation methodology

3.1 Method selected

We conducted:

1. **Think-aloud protocol**, and
2. **Usability testing with observations**.

No heuristic or expert review was used. All insights were derived directly from user behaviour, verbalizations, and moderator observations.

3.2 Rationale

- Think-aloud reveals user mental models, confusion points, and expectations in real time.
- Usability testing provides systematic task-based performance data rooted in actual user behavior.
- This method fits the goal of assessing interpretability and UI comprehension because participants express what they understand while interacting with the system.

3.3 Procedure

Participants: Five participants were selected.

Tasks:

1. Fill the input form for a sample employee.
2. Interpret the predicted absenteeism result.
3. Examine local explanations (SHAP/LIME).
4. Explore "what-if" cases using counterfactual explanations.

Session format: Participants were asked to verbalize all thoughts while performing tasks. The moderator observed behaviours, captured comments, and recorded interaction difficulties.

4 (c) Findings

4.1 Comprehension and labeling

- Multiple participants struggled with label meanings (e.g., "Hit Target", "Temporal Factor").

- Units were unclear (e.g., Workload unit, Units per Day).

4.2 Explainability comprehension

- Users did not know what SHAP and LIME were without explanation.
- When shown top contributing features, most users could understand the direction and impact of features.

4.3 Layout and visual design

- Full-screen width not utilized; empty margins reduced perceived efficiency.
- Some information hidden in collapsible sections was overlooked entirely.

4.4 Performance and reliability

- Slowness in explanation generation reduced trust in system responsiveness.

4.5 Fairness & trust

- Poor model performance metrics caused hesitation in relying on predictions.
- Users did not understand how to interpret fairness indicators without guidance.

4.6 Representative anonymized quotes

- "Why is 'Hit Target' under Family Information?"
- "What is LIME or SHAP supposed to mean?"
- "This screen has a lot of empty space—use it properly."
- "Season feature is unnecessary; add more absence reasons instead."

5 (d) Improvements

5.1 UI-level

1. Clarify labels and move confusing fields to appropriate sections.
2. Add tooltips and unit explanations next to numerical inputs.
3. Expand layout to make use of full width.
4. Add brief explanatory text next to SHAP/LIME outputs.

5.2 Explainability

1. Provide small text definitions for SHAP, LIME, and counterfactuals.

5.3 Performance

1. Compute heavy explanations asynchronously and show progress indicators.
2. Review caching strategy for global explanations to maintain responsiveness.

6 (e) Reflections on method

6.1 What worked well

- Think-aloud produced rich insights into user misunderstandings.
- Observation notes captured consistent patterns across users.

6.2 Limitations

- Small sample size limits generalizability.
- Lack of expert review means some design issues may remain undetected.
- No quantitative metrics (SUS, comprehension scores) were collected.

6.3 Future improvements

- Increase participant diversity.
- Add SUS + comprehension questionnaire.
- Record and transcribe sessions for deeper qualitative analysis.

7 (f) Evidence

Links:

- **Observation Notes:** https://drive.google.com/file/d/1sLR9whSzmnp5RJuFv2h2PUhjCXVcdmcb/viev
- **Recording Folder:** https://drive.google.com/drive/folders/11bGoKdJ9vPKhA$_5 7r1CxpUbA9nZ98tjb$
- **Survey Response Sheet:** https://docs.google.com/spreadsheets/d/120gKAaHYc-4RnAsj8Fab66-Zqrp0GTH0/edit

8 Conclusion

The think-aloud and usability-based evaluation revealed that labeling, clarity of units, and discoverability of explanations require improvement. Trust-related issues stem from model performance. The recommended improvements directly address these issues and enhance both usability and explainability.