Final Report

Interactive and Explainable AI

Stijn But Minji Kim Xuechun Lyu Sercan Şeref

2025-05-27

This report presents the design, development, and evaluation of an interactive explainable AI (XAI) dashboard aimed at helping data science students—particularly first-time home buyers—interpret housing price predictions. Addressing the "disagreement problem" in XAI, the dashboard enables users to compare multiple explanation methods (SHAP, LIME, Integrated Gradients, SmoothGrad, and GradientShap) side-by-side, fostering transparency and trust in model outputs. The project combines user research, creative ideation, and iterative prototyping to deliver a tool that balances technical rigor with usability. Qualitative user testing highlights the dashboard's effectiveness in clarifying feature importance and model behavior, while also identifying areas for further improvement in interpretability and user experience. The findings contribute practical insights for designing accessible XAI tools in real-world decision-making contexts.

Table of contents

1 Introduction

The disagreement problem, as highlighted by Kaur, emphasizes the challenge of interpreting conflicting feature importances provided by different explanation tools and methods. (Kaur et al. 2020) This issue is particularly relevant in the context of dashboarding and explanation tools, where users often struggle with varying explanation styles and visualizations. The lack of a standardized approach to feature importance can lead to confusion and misinterpretation, especially for users who may not have a deep understanding of the underlying data science concepts.

Our project addresses this challenge by designing a dashboard specifically tailored for data science students. The dashboard aims to help users understand the features that contribute to housing price predictions by enabling them to compare multiple explanation methods side-by-side. This comparison provides insights into how different methods attribute importance and supports users in interpreting these explanations more effectively.

The relevance of such a tool is particularly strong when considering the Dutch housing market, where housing prices have risen sharply in recent years. According to CBS, the prices of existing homes are now higher than during the previous peak in 2008, with the pace of price increases slowing slightly around 2019 before accelerating again. (Statistiek 2024) In a market where finding affordable housing is increasingly challenging, a tool that explains housing price predictions in an accessible and transparent way can help users better understand the factors driving property prices and support more informed decision-making.

We chose to focus on first-time house buyers, as they often face difficulties in understanding which features contribute most to a home's value. Given that first-time buyers are usually early in their careers and lack prior investment experience, they stand to benefit greatly from clear and interpretable AI explanations. Targeting data science students within this group was a deliberate choice, as their foundational knowledge of data concepts allows them to engage with and benefit from the explanations provided by the dashboard more effectively.

- *Design debrief*

To achieve this, we conducted user research and pilot testing using standard dashboarding tools. Based on the findings, we designed a prototype dashboard that aligns with the needs of our target audience. The datasets used for this project are sourced from the course materials or other relevant projects.

2 Empathize

2.1 Explanation Methods in Machine Learning

We utilized various explanation methods introduced during the initial lectures and explored through the notebooks provided by the lecturers, including SHAP, LIME, Integrated Gradients, SmoothGrad, and GradientShap. These notebooks not only helped us understand the theoretical aspects of these methods but also demonstrated their practical applicability, showing that we could effectively use these techniques to address the housing problem in the Netherlands. Each of these methods offers unique approaches to understanding feature importance and model interpretability.

The emergence of Explainable Artificial Intelligence (XAI) represents a critical development in addressing the opacity of complex machine learning models. Traditional predictive models, particularly those employed in high-stakes domains such as finance, healthcare, and housing economics, often suffer from a lack of interpretability. To bridge this gap, a variety of explanation techniques have been proposed, each offering different perspectives on how input features contribute to model outputs.

Local Interpretable Model-Agnostic Explanations (LIME), introduced by Ribeiro et al. (2016), is a seminal contribution in this regard. LIME operates by approximating a complex model locally around a prediction using a simpler, interpretable surrogate model, often a linear regression. Through perturbing input data and observing output variations, LIME offers intuitive explanations that are particularly useful in understanding the behavior of highly non-linear models.

Another important advancement is SHapley Additive exPlanations (SHAP), formulated by Lundberg and Lee (2017). Rooted in cooperative game theory, SHAP assigns each feature an importance value for a particular prediction by considering the contribution of features across all possible combinations. SHAP stands out due to its axiomatic foundation, guaranteeing properties such as local accuracy, consistency, and missingness, which are crucial for ensuring credible model interpretations.

Integrated Gradients, proposed by Sundararajan et al. (2017), takes a different approach, specifically designed for interpreting deep neural networks. This method attributes the change in prediction between a baseline and the actual input by integrating the gradients along a linear path. It satisfies important theoretical properties, such as sensitivity and implementation invariance, making it particularly suited for continuous and complex input spaces like images or tabular financial data.

Another notable method is SmoothGrad, introduced by Smilkov et al. (2017), which improves the clarity of saliency maps by adding noise to the inputs and averaging the resulting gradients. Although initially proposed for visual data, adaptations of SmoothGrad to tabular data offer enhanced feature visualization by reducing noise and highlighting the regions of true importance.

Overall, these explanation methods each bring unique strengths. LIME offers model-agnostic, localized explanations ideal for exploratory analysis; SHAP provides a globally consistent, theoretically sound framework; Integrated Gradients excel in deep learning contexts; and SmoothGrad enhances the robustness and visual clarity of explanations. The synergy of these techniques creates a comprehensive interpretability toolkit essential for advancing transparent and trustworthy machine learning applications in various domains, including the housing market.

2.2 Applications of Explanation Methods to Housing Market Analysis

The housing market has historically been analyzed through hedonic pricing models, wherein property characteristics such as location, size, and amenities are linked to price. However, with the advent of machine learning, more sophisticated models like Random Forests, XGBoost, and deep neural networks have demonstrated superior predictive capabilities. These advancements, while improving accuracy, have exacerbated concerns about model transparency, particularly in socially and economically sensitive sectors such as real estate.

The application of XAI methods to housing market analysis addresses this issue by elucidating the underlying drivers of model predictions. For instance, Özçelik and Yildirim (2022) conducted a comparative study applying SHAP and LIME to real estate valuation models. Their findings consistently demonstrated that variables such as location proximity to urban centers, size of the dwelling, quality of neighborhood amenities, and macroeconomic indicators such as interest rates are the dominant predictors of property prices. Importantly, SHAP and LIME provided granular, instance-specific insights that enabled a deeper understanding of the multifaceted factors influencing real estate valuation.

Feature importance analyses using SHAP and LIME have revealed recurrent patterns across various studies. Location factors, such as distance to city centers and accessibility to public transport, emerge as primary determinants of housing prices. Demographic variables, including median income levels and employment rates, also exhibit significant influence. Moreover, market dynamics such as housing supply-demand ratios and mortgage interest rates are critical economic indicators that affect property values. Physical attributes of houses, including size, number of rooms, age, and the presence of amenities such as gardens or parking spaces, consistently appear among the top predictors across diverse datasets.

A particularly novel contribution to this field is the work by De Nadai et al. (2016), who utilized mobile phone activity data to quantify urban vitality, subsequently demonstrating its predictive power for housing price fluctuations. This study highlighted the potential of integrating unconventional datasets and features into traditional housing models, further underscoring the versatility of XAI methods in uncovering hidden patterns.

Focusing specifically on the Dutch housing market, reports by Statistics Netherlands (CBS) and research conducted by Rabobank indicate distinct patterns that are critical for modeling efforts. Urbanization has driven significant price increases within the Randstad metropolitan region compared to rural provinces. ((CBS) 2024) Fluctuations in mortgage interest rates have shown a strong correlation with transaction volumes, emphasizing the sensitivity of the housing market to macroeconomic policy changes. Additionally, government interventions such as rent control policies and adjustments in mortgage lending standards have significantly influenced market dynamics. (Research 2024)

In constructing educational dashboards aimed at data science students, the integration of explanation methods is particularly advantageous. Visual tools such as SHAP summary plots and LIME explanation graphs allow users to intuitively grasp the complex interplay of features driving housing market predictions. Furthermore, scenario simulation functionalities, wherein users can modify input features and observe corresponding changes in predictions, offer a hands-on understanding of model behavior. According to Molnar (2020), effective communication of explanations requires careful consideration of the audience's domain knowledge, making simplicity, visual clarity, and contextual relevance crucial design principles.

In conclusion, the integration of explanation methods into housing market analysis not only enhances model transparency but also facilitates a deeper comprehension of the economic, demographic, and physical factors shaping real estate dynamics. This synergy between advanced predictive modeling and interpretability is particularly valuable in educational contexts, equipping data science students with the necessary skills to build, critique, and trust predictive systems deployed in real-world scenarios.

2.3 User Research and Pilot Testing

To gain a deep and empathetic understanding of the problem space and our potential users, we conducted exploratory user research centered around a prototype dashboard for explainable AI in the housing domain. Although the dashboard was not yet fully deployed in a real-world setting, participants engaged with interactive mockups and guided walkthroughs that closely simulated the intended user experience. This approach enabled us to evaluate early design concepts and explanation strategies using qualitative methods, including pilot testing, semi-structured interviews, and observational feedback.

Our target users were selected for their foundational knowledge of AI and predictive modeling, making them well-suited to critically assess the interpretability and usability of a broad range of XAI methods. This group was intentionally chosen over a broader young adult audience because, while young people in general are a key demographic for first-time home buying, their typically limited experience with data science would have required the dashboard to be extremely simplistic. Achieving such simplicity is particularly challenging with advanced explanation tools such as SHAP, LIME, Integrated Gradients, SmoothGrad, and GradientShap, all of which inherently involve complex visualizations and concepts.

By focusing on data science students, we could design a dashboard that leverages users' existing familiarity with data-driven reasoning, allowing for more nuanced explanations and richer interactions across multiple explanation techniques without overwhelming the audience. Each participant took part in a semi-structured session, guided by a structured feedback matrix (covering Likes, Criticisms, Questions, and Ideas). During these sessions, participants explored how the dashboard visualized housing price predictions using a variety of explanation methods—including global (e.g., SHAP) and local (e.g., LIME) model-agnostic techniques, as well as gradient-based approaches (Integrated Gradients, SmoothGrad, GradientShap) for neural networks. This format encouraged participants to reflect on the clarity, usefulness, and trustworthiness of the outputs, share their preferences, and articulate any difficulties or suggestions for improvement.

Several recurring insights emerged from this process. Visual clarity and intuitive color schemes were repeatedly highlighted as essential, with SHAP and gradient-based methods often praised for their structured visuals and consistent layout. In contrast, some explanation outputs—especially those from LIME and certain neural network methods—were perceived as fragmented or more difficult to interpret without additional guidance. The inclusion of a side-by-side explanation comparison was particularly valued, as it made disagreements between methods transparent and encouraged critical reflection on model behavior.

However, the exploratory nature of the dashboard also revealed certain limitations. Some users expressed confusion over changing feature importances between samples and across explanation methods, raising concerns about the consistency and reliability of explanations. Others found abstract indexing and the absence of domain-specific context (such as geographic information or real housing listings) to be barriers to understanding. These findings underscored the importance of contextualization and the need for onboarding materials, even in early-stage XAI tools.

Despite being a conceptual prototype, the dashboard functioned as a research probe, surfacing how users think about trust, explanation alignment, and AI transparency across a variety of XAI techniques. The pilot testing confirmed that even at an early stage, well-designed visualizations and interactive comparisons can elicit valuable user reactions, inform future design decisions, and align with real user needs. Ultimately, these insights grounded our assumptions in authentic user experience and highlighted key priorities for future development and deployment in fully functional, real-world systems.

3 Define

Leading into the ideation phase, we consolidated insights from both our technical exploration of explainability techniques and the qualitative feedback gathered during early user engagement. This synthesis helped us clearly articulate the problem space and select the most appropriate XAI methods for supporting interpretability in the context of housing price prediction.

Through our interviews and pilot testing with data science students, two primary user needs emerged. First, users expressed a desire to understand the role of input features at both a general and individual level. They wanted to grasp how features influence predictions across the entire dataset—such as the general importance of property size or location—as well as understand how those same features contribute to specific predictions for individual houses. Second, many users struggled when faced with conflicting explanations from different XAI methods. This issue, closely aligned with what Kaur et al. (2020) describe as the “disagreement problem,” led to uncertainty and reduced trust in the model’s outputs. Users clearly needed support in interpreting and reconciling these differing perspectives.

Our initial scope included a broad range of explanation methods: SHAP, LIME, Integrated Gradients, SmoothGrad, and GradientShap. This selection was motivated by course material, literature, and the technical affordances of each method. However, early user testing revealed key limitations. Gradient-based methods, while theoretically compelling and powerful in neural network contexts, produced explanations that were difficult to interpret within our Streamlit-based interface. The visual outputs of techniques like Integrated Gradients and SmoothGrad—typically saliency maps or noisy bar charts—proved less accessible for users unfamiliar with deep learning or advanced visualization formats.

In contrast, SHAP and LIME consistently resonated with users. SHAP’s strength lies in its ability to provide both global and local explanations with a solid theoretical foundation rooted in Shapley values. Users responded positively to the visual clarity of SHAP’s waterfall and summary plots, particularly when these linked feature contributions to monetary impacts, making it highly relevant in the housing domain. LIME offered a complementary perspective. While it required a greater cognitive effort to interpret, its case-specific surrogate models enabled users to explore how slight changes in inputs could lead to different predictions—adding depth to the interpretability experience.

Focusing our prototype on SHAP and LIME allowed us to reduce unnecessary complexity, streamline the interface, and ensure that explanation outputs remained accessible to data science students. It also enabled more meaningful side-by-side comparisons, which directly addressed users’ confusion around explanation inconsistency. In short, this narrowing of scope not only improved usability but sharpened the pedagogical value of the dashboard.

This led us to a well-defined design challenge: how might we help data science students critically compare and interpret different explanation methods for housing price predictions—especially when those methods disagree? With this framing in mind, the ideation phase focused not just on creating an explainability dashboard, but on designing a comparative explanation interface. The goal was to promote reflective use, guide interpretation, and foster trust in machine learning models by highlighting divergence, offering contextual support, and encouraging informed judgment.

4 Ideation

4.1 Description of the creative techniques used for divergence and convergence

The ideation phase began after we had developed a clear understanding of both the technical landscape of explainable AI (XAI) methods and the cognitive and emotional needs of our target audience: data science students interested in understanding housing price predictions. At this point, our task was not to solve the problem immediately, but to explore its full creative potential through structured divergence and convergence.

We initiated the divergence process with a brainstorming session framed by the question: **"How can we make explanation method results easier to understand for people?"** This was designed to be as inclusive and expansive as possible, welcoming both practical and speculative ideas. To support creativity and collaboration, we applied the "Yes, and..." technique. This ensured that no idea was dismissed too early and that suggestions could grow organically through group interaction.

The outcomes of this phase reflected a wide range of thinking. Some ideas focused on clarity and visual communication—for example, using red and green color coding to indicate negative and positive feature impacts. Others emphasized personalization and engagement, such as offering simplified or advanced visualizations based on the user's experience level. Some were clearly ambitious, such as an AI-powered estate agent capable of responding in natural language, or a "What if?" simulator allowing users to tweak house features and view resulting price predictions. While not all of these ideas would be implemented, their diversity reflected a strong understanding of both user needs and the explanatory potential of different XAI techniques.

After this expansive creative exploration, we began the convergence phase. Here, we applied the COCD Box method—a structured decision-making tool that categorizes ideas by their feasibility and innovativeness. Each idea was placed into one of four zones: Blue (feasible and easy to implement), Red (innovative and easy to implement), Yellow (innovative but harder to implement), and Grey (expensive or complex to implement). This helped the team identify which ideas could be prioritized for prototyping and which needed to be discarded or saved for future development.

Ideas such as using SHAP and LIME together, adding clear labels, using intuitive visual encodings (like SHAP waterfall plots), and including a feature comparison table landed in the Blue Zone. These were feasible, technically grounded, and aligned directly with user needs. In the Red Zone, we placed enhancements like emojis and animations that would humanize the dashboard without overwhelming users. While the Yellow and Grey Zones held intriguing concepts—such as integrating gradient-based methods for neural networks or implementing an AI estate agent—we ultimately set them aside due to limitations in the Streamlit framework and interpretability concerns during pilot testing.

This creative process—from wild exploration to structured selection—ensured that we didn't prematurely settle on an idea and that the final solution reflected both innovation and usability, grounded in direct user feedback and technical viability.

4.2 Description of the chosen solution

The chosen solution that emerged from our ideation phase is a comparative explanation dashboard designed specifically for data science students seeking to understand housing price predictions. The core aim of this solution is to resolve the "disagreement problem" described in Kaur et al. (2020) by enabling users to interpret and critically compare the outputs of two widely used explanation methods: SHAP and LIME. These methods were selected not only for their technical relevance, but also for their complementary strengths in supporting both global and local interpretability.

Rather than overwhelming users with a wide array of explanation methods, the dashboard deliberately focuses on SHAP and LIME to keep the interface intuitive while supporting meaningful comparison. SHAP, which is based on Shapley values, offers strong theoretical grounding and was chosen for its dual capability to provide both global explanations (via summary plots) and local explanations (via waterfall plots). Feedback from early prototyping confirmed its usability and appeal, especially when the feature attributions were expressed in monetary terms. LIME was chosen to complement SHAP's global focus by offering localized surrogate model explanations that show how small changes in input affect individual predictions.

While the initial scope included gradient-based methods such as Integrated Gradients and SmoothGrad, these were excluded from the final interface due to performance limitations and interpretability challenges. Users found their visual outputs less accessible, especially when rendered through Streamlit's limited plotting options. This decision reflected our design principle of prioritizing clarity over complexity.

The solution was thus refined into a side-by-side explanation interface that allows users to select a housing sample, view SHAP and LIME explanations for the same prediction, and compare the top features identified by each method. This comparison is supported by an overlap analysis, prompting users to reflect on where explanations agree or diverge. These design decisions were made to promote transparency, reduce cognitive load, and foster interpretability and trust in machine learning predictions—especially in the high-stakes context of housing.

5 Prototype

The chosen solution that emerged from the ideation phase is an interactive, explanation-focused dashboard that enables data science students to compare SHAP and LIME outputs side-by-side in the context of housing price predictions. This design directly addresses the "disagreement problem" by making explanation variability visible and interpretable, thus supporting user reflection and fostering trust in AI-driven decision-making.

Built using Streamlit, the dashboard operationalizes this concept through a clean interface supported by a robust modeling pipeline. The underlying dataset, sourced from Kaggle and focused on the Dutch housing market, undergoes thorough preprocessing—including outlier removal via z-score filtering, one-hot encoding of categorical variables, and normalization of features—to ensure consistent model behavior.

Initially, the architecture supported two models: an XGBoost regressor and a feedforward neural network. However, after testing revealed usability limitations and interpretability challenges associated with gradient-based neural explanations (Integrated Gradients, SmoothGrad, and GradientSHAP), the neural network was removed from the final prototype. This decision helped streamline the dashboard and align it more closely with user needs, focusing on the more accessible and pedagogically effective XGBoost model.

The SHAP explanation module presents both local and global insights. A waterfall plot visualizes how each feature pushed the selected house's prediction above or below the model's baseline, while a summary plot aggregates feature importance across the full test set. Both visualizations benefit from color coding and integrated tooltips, making them approachable for novice users while still providing sufficient depth for advanced learners.

In parallel, the LIME explanation module offers a case-specific analysis by generating a local surrogate model around the selected prediction. Rendered in an embedded HTML panel, the LIME output highlights feature intervals and their respective contributions. Though initially less intuitive for users, iterative refinements such as improved tooltip language and clearer labels helped make LIME more interpretable.

One of the dashboard's most distinctive features is the feature comparison module, which identifies and displays the top five features according to SHAP and LIME for the same prediction. The module calculates overlap between the methods and presents this visually, prompting users to engage with questions such as: Why do these methods disagree? What can I learn from that? This element serves as a direct response to the interpretability challenge at the heart of the project.

To further support contextual understanding, a feature overview panel displays the original (non-normalized) values of input features for the selected house. This helps bridge the gap between abstract model explanations and real-world property characteristics. Additionally, hoverable tooltips are embedded throughout the dashboard to explain key terms and graph components in accessible language.

User feedback is collected through an interactive sidebar, where participants can report which explanation method they found easier to understand and more trustworthy. An open comment box invites further suggestions, enabling an iterative development process grounded in participatory design.

Ultimately, this solution transforms theoretical XAI concepts into a functional educational tool. It empowers users to not only view explanations but also critically compare them, reflect on their reliability, and contextualize model predictions in a meaningful way. The dashboard doesn't just present results—it facilitates understanding. Through careful integration of accessible design, model transparency, and user reflection, it fulfills the project's central design challenge: to make machine learning explanations interpretable, comparable, and actionable for its intended audience.

6 Test

6.1 Qualitative User Research Methods

To evaluate the effectiveness of our dashboard in addressing the disagreement problem and supporting interpretability, we conducted qualitative user research with our target audience: master's students in data science, AI, and computer science. Our evaluation strategy combined structured user testing with semi-structured feedback collection. This approach ensured we could assess both usability and conceptual clarity from multiple perspectives.

Each user participated in a guided test session, during which they freely interacted with the dashboard. Tasks included selecting housing samples, exploring SHAP and LIME explanations, and interpreting the comparative feature analysis. We provided minimal guidance during the session to observe natural engagement and confusion points. These sessions were supplemented by a structured feedback table that captured method preference, visual design opinions, criticisms, questions, and improvement suggestions.

6.2 Research Question

The central research question guiding our user testing was:

- **"How do data science students perceive and interpret the explanations provided by different XAI methods in the dashboard, and what are their preferences, challenges, and suggestions for improving the interpretability and usability of housing price prediction models?"**

This question aimed to investigate whether the prototype could effectively support users in understanding the behavior of complex models and resolving discrepancies between different explanation outputs.

6.3 Results of the Data Analysis

Analysis of user responses revealed several important trends. First, SHAP was the overwhelmingly preferred explanation method, praised for its clarity, intuitive visuals, and direct representation of feature influence (e.g., monetary values). Users described SHAP as easier to interpret and more informative, especially when comparing feature contributions.

In contrast, LIME received mixed feedback. While some users found value in its localized explanations, others struggled with redundant or unclear visualizations. Comments frequently mentioned "too many graphs," unclear terminology (such as "sample index"), and cognitive overload. There was a clear demand for better onboarding, including a brief user guide, tooltips, or an introductory tutorial explaining how each explanation method works and how to interpret the outputs.

A recurring issue was confusion over the lack of overlap between explanation methods. Users often questioned which method to trust when SHAP, LIME, and gradient-based explanations disagreed. This confirmed the relevance of the disagreement problem (Kaur et al., 2020) and underscored the need for improved explanation harmonization or contextual guidance.

Several usability issues were also identified. Participants noted slow dashboard performance, especially when switching between complex models like neural networks. Others found the layout of some visualizations (particularly LIME) overwhelming, and recommended simplifying visual elements or providing clearer sectioning. Suggestions included replacing generic index labels with more meaningful identifiers (e.g., "house #12"), adding personalization, and improving visual design polish.

An additional challenge emerged from the dashboard's dual-model structure: both the XGBoost and Neural Network tabs provided SHAP and LIME explanations, while the Neural Network tab also included gradient-based methods—Integrated Gradients, SmoothGrad (via NoiseTunnel + IG), and GradientSHAP. Users frequently struggled to interpret the differences between SHAP and LIME explanations across the two model tabs, and found it difficult to understand how the neural network-specific methods related to the more familiar XGBoost explanations. During user testing, it was also challenging for the team to concisely explain these distinctions, which sometimes led to confusion and reduced confidence in the outputs.

From this research, we derived the following key insights for improvement: users need a concise introduction or guide to help them navigate the dashboard and understand explanation methods; visual clutter, especially in LIME sections, should be reduced, with more intuitive labels and polished layouts; replacing index numbers with meaningful names (such as "House A" or "House in Amsterdam") would improve comprehension; and faster model switching and reduced load times would enhance the fluidity of user interaction. Additionally, clearer communication about the differences between model types and their associated explanation methods is needed to help users make sense of the outputs.

In summary, the user research provided valuable insights into how moderately technical users engage with XAI tools and what barriers exist in understanding model behavior. The results confirmed the dashboard's overall value, especially in enabling comparison and transparency, but also revealed critical usability and communication gaps. These insights are now guiding further iterations—focusing on simplifying LIME outputs, integrating layered explanations, clarifying model and method differences, and improving interaction design to enhance overall clarity and trust in AI predictions.

7 Conclusion and Recommedations

7.1 Conclusion of the User Testing

The user testing phase validated the core concept of the dashboard and confirmed its utility in addressing the "disagreement problem" in explainable AI. By enabling side-by-side comparisons of SHAP and LIME outputs, the dashboard made explanation differences transparent and promoted user reflection on model reasoning. Our participants found the interface intuitive and informative, especially in understanding which features influenced housing price predictions.

SHAP was widely favored for its structured and visually intuitive representation of feature contributions. The inclusion of monetary values helped contextualize the predictions, making them easier to interpret. LIME was appreciated for its case-specific explanations but drew criticism for being more difficult to interpret, especially due to visual clutter and less intuitive output. Despite these challenges, users valued having both methods present, as this encouraged critical comparison and deeper engagement with model behavior.

Several constructive suggestions emerged from testing. Users requested better onboarding, including concise tooltips or an introductory guide—particularly to support interpretation of LIME. Others recommended improving sample labeling by replacing abstract index numbers. A few participants also noted that switching between models could cause noticeable performance lags.

Useres also found it difficult to compare explanations across tabs. Because toggling between the XGBoost and neural network tabs required waiting for the dashboard to reload, users could not easily keep previous outputs in mind. This made it nearly impossible to spot differences between the explanations provided by each model, further complicating interpretation and reducing the effectiveness of side-by-side comparison.

These insights informed several important changes in the prototype. First, we removed the neural network tab entirely. Although it featured additional explanation methods (e.g., GradientSHAP, SmoothGrad via NoiseTunnel, and Integrated Gradients), their outputs were difficult to interpret due to Streamlit's limited native support for more advanced visualizations. The neural network section also introduced new complexity, sparking additional user questions about how its outputs differed from those of models like XGBoost. In light of the limited clarity and increased cognitive load, removing this tab helped streamline the overall user experience.

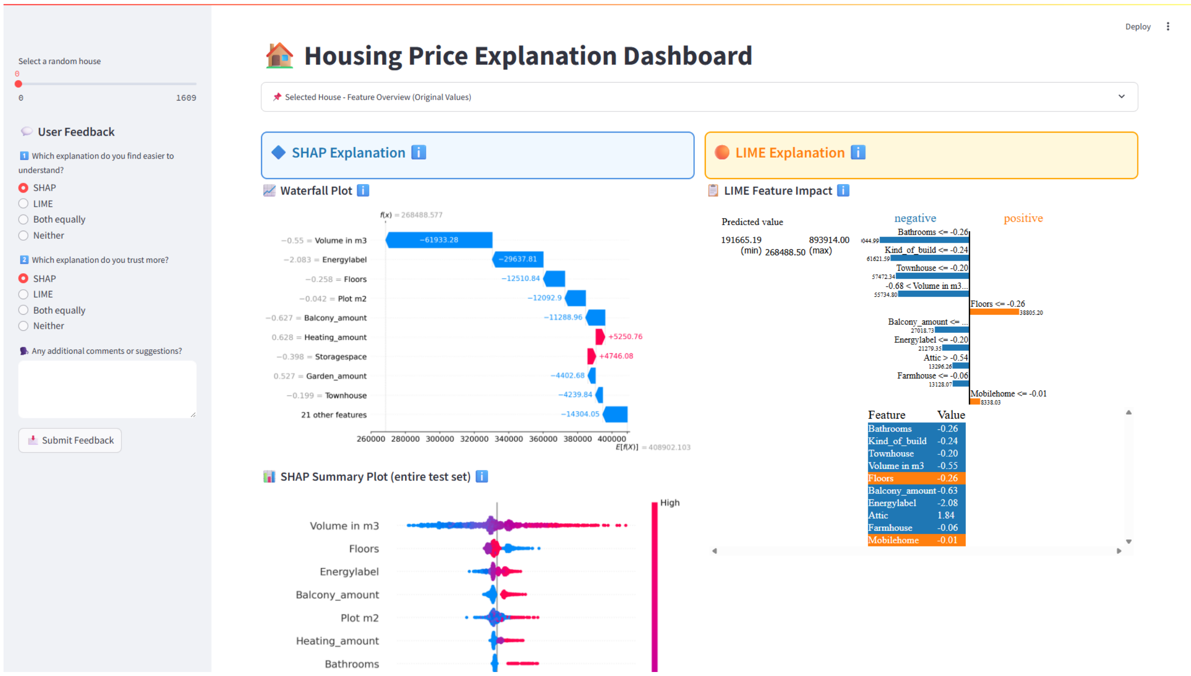
Additionally, we revised several tooltips to improve clarity and approachability, aligning them better with the data science students' level of familiarity. We also refined interface terminology—for instance, updating labels like "Select a random index" to more understandable phrases like "Select a random house." This change improved user orientation and made the dashboard feel more aligned with its real-world use case.

One of the most impactful improvements was the addition of a toggle that allows users to view the original value of each feature. This feature helps place SHAP and LIME outputs in context by showing not only how much a feature influenced the prediction, but also what specific value it had. This addition significantly enhanced interpretability, allowing users to trace how real-world inputs—like a home's square footage or number of rooms—translated into changes in estimated price.

Together, these refinements enhanced the dashboard's clarity, usability, and alignment with user expectations. The result is a tool that not only fosters trust in AI predictions but also promotes meaningful user engagement with model explanations in a high-stakes decision-making context like real estate.

7.2 Description of the Final Prototype with Visualizations

The final prototype expands on the core structure of the initial implementation, offering a complete, polished dashboard that enables users to explore, compare, and contextualize AI predictions in an interactive and educational environment. Designed with feedback from multiple user testing sessions, the final version maintains the dual-method focus on SHAP and LIME while enhancing usability, interactivity, and clarity.



Select a random house

0

0 1609

User Feedback

1️⃣ Which explanation do you find easier to understand?

● SHAP
○ LIME
○ Both equally
○ Neither

2️⃣ Which explanation do you trust more?

● SHAP
○ LIME
○ Both equally
○ Neither

Any additional comments or suggestions?

Submit Feedback

🏠 Housing Price Explanation Dashboard

📌 Selected House - Feature Overview (Original Values)

◆ SHAP Explanation ℹ

📈 Waterfall Plot ℹ

📊 SHAP Summary Plot (entire test set) ℹ

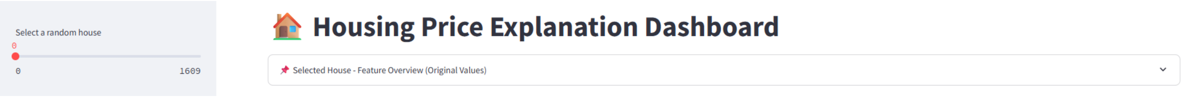
● LIME Explanation ℹ

📋 LIME Feature Impact ℹ

Feature	Value
Bathrooms	-0.26
Kind_of_build	-0.24
Townhouse	-0.20
Volume in m3	-0.55
Floors	-0.26
Balcony_amount	-0.63
Energylabel	-2.08
Attic	1.84
Farmhouse	-0.06
Mobilehome	-0.01

Figure 1: Dashboard Overview

Upon launching the dashboard, users are welcomed by a streamlined interface that begins with the selection of a random house from the dataset. This selection dynamically updates the dashboard, triggering the rendering of explanations tailored to the chosen property. An expandable panel provides the original, non-normalized feature values, anchoring abstract explanations in real-world data.



Select a random house

0

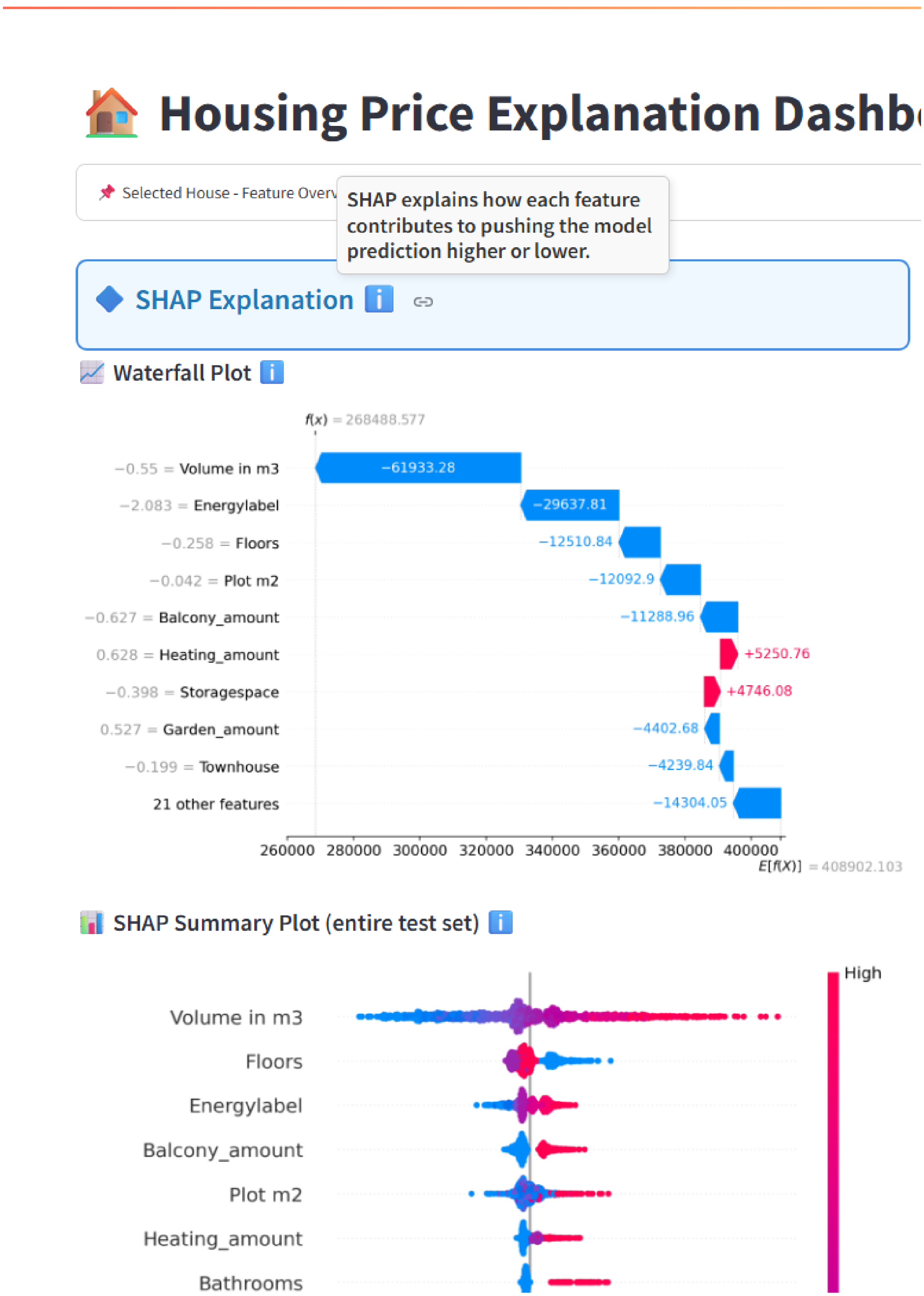
0 1609

🏠 Housing Price Explanation Dashboard

📌 Selected House - Feature Overview (Original Values)

Figure 2: Random House Selection and Expendable Panel of Original Feautures

The explanation interface is organized into two symmetrical panels. On the left, the SHAP module includes a waterfall plot that visualizes how each feature pushes the prediction higher or lower for the selected sample. Below it, a summary plot presents global feature importance across the dataset, with color-coding (red for high values, blue for low) helping users intuitively grasp trends. Hoverable tooltips provide immediate, simple explanations for each plot element, aiding interpretation.



Housing Price Explanation Dashb

📌 Selected House - Feature Overv

SHAP explains how each feature contributes to pushing the model prediction higher or lower.

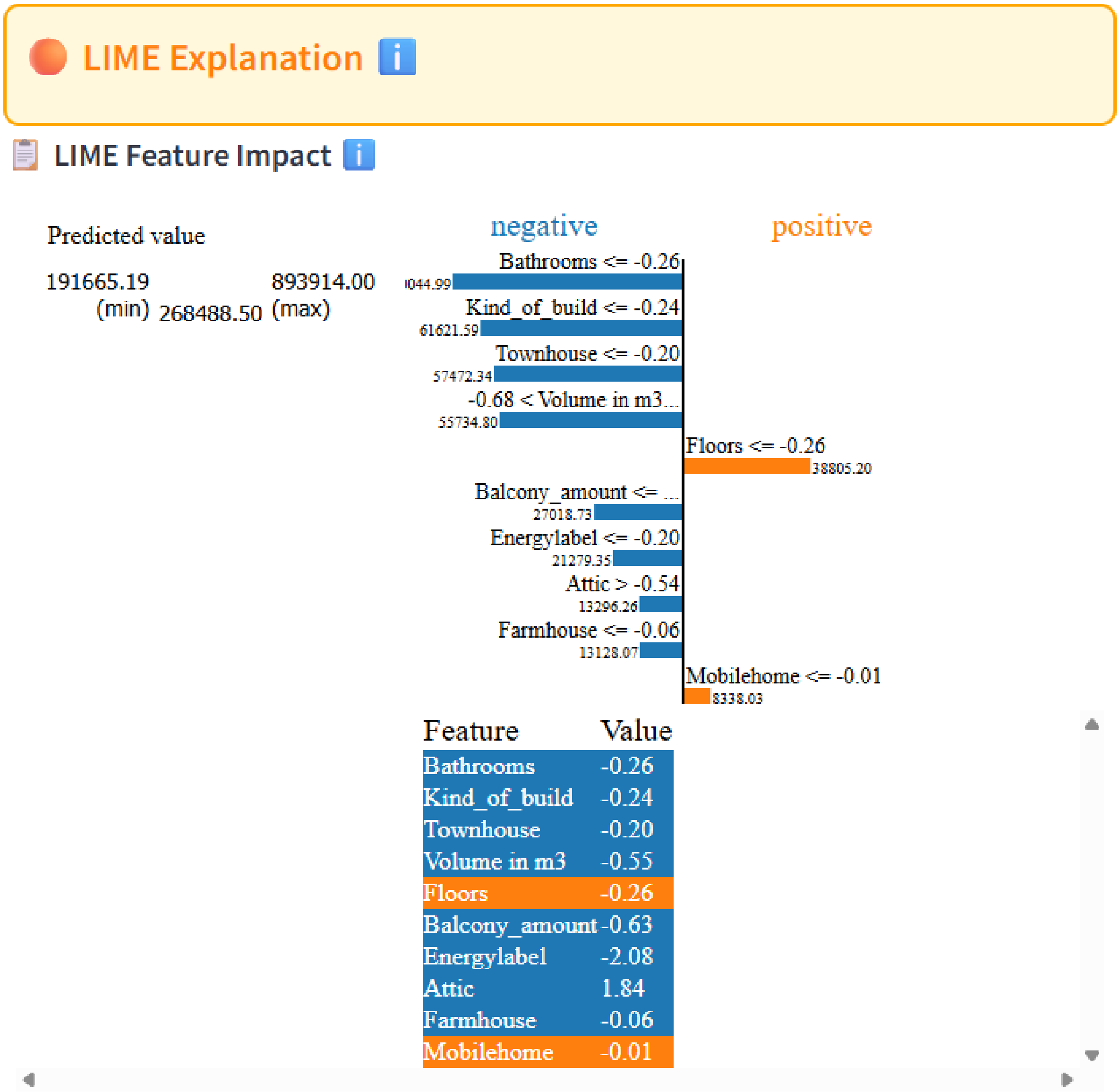
SHAP Explanation

Waterfall Plot

SHAP Summary Plot (entire test set)

Figure 3: SHAP Waterfall and Summary Plots

On the right side, the LIME module delivers a local explanation for the same house using an HTML-rendered visualization. This shows how individual feature conditions contributed to the prediction, based on slight perturbations to the inputs. Tooltips again support user understanding by clarifying the method's logic and the meaning of the visual outputs.



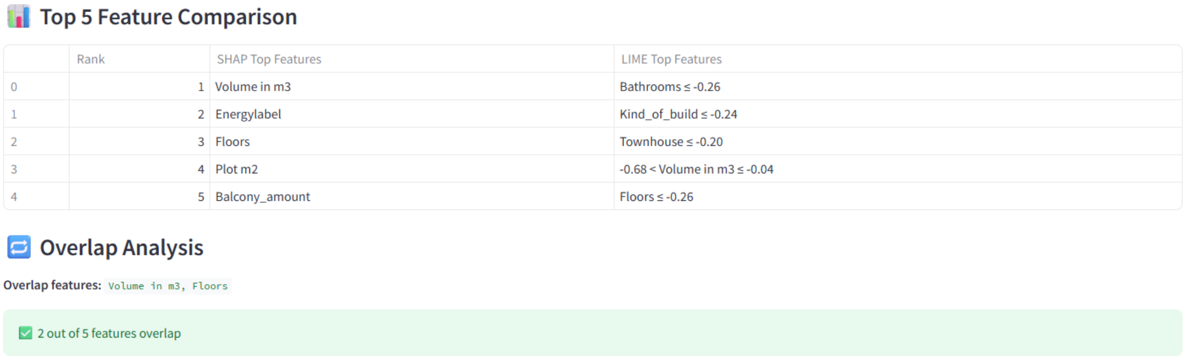
LIME Explanation

LIME Feature Impact

Feature	Value
Bathrooms	-0.26
Kind_of_build	-0.24
Townhouse	-0.20
Volume in m3	-0.55
Floors	-0.26
Balcony_amount	-0.63
Energylabel	-2.08
Attic	1.84
Farmhouse	-0.06
Mobilehome	-0.01

Figure 4: LIME Explanation Plots

A central innovation is the feature comparison module. This component extracts the top five features highlighted by SHAP and LIME for the selected prediction, displays them in a structured table, and highlights any overlap. By explicitly drawing attention to agreement or disagreement, the dashboard fosters deeper user engagement with the reasoning processes behind AI predictions and encourages reflection on why models may disagree.



Top 5 Feature Comparison

	Rank	SHAP Top Features	LIME Top Features
0	1	Volume in m3	Bathrooms ≤ -0.26
1	2	Energylabel	Kind_of_build ≤ -0.24
2	3	Floors	Townhouse ≤ -0.20
3	4	Plot m2	-0.68 < Volume in m3 ≤ -0.04
4	5	Balcony_amount	Floors ≤ -0.26

Overlap Analysis

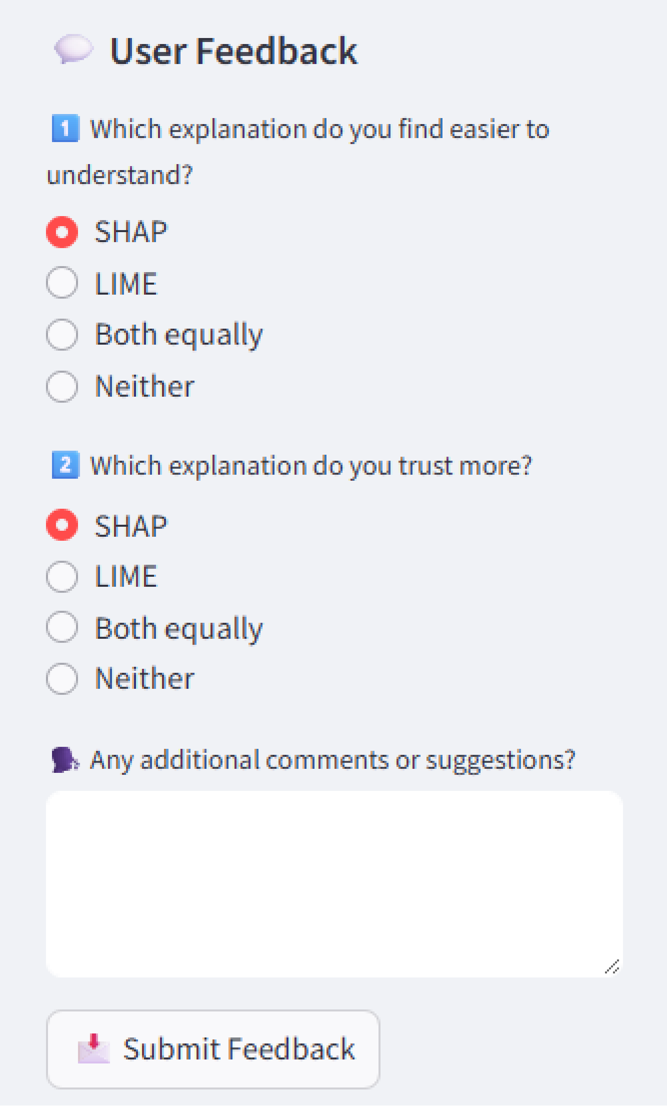
Overlap features: `Volume in m3, Floors`

☑ 2 out of 5 features overlap

Figure 5: Top Feature Comparison Between SHAP and LIME

Another important feature added during development was a toggle to reveal the original value of each feature for the selected house. This contextual view allows users to see not just that a feature influenced a prediction, but how its specific value (e.g., "Living space: 140 m^2") contributed to the outcome. Labels were also revised to replace abstract numerical indices with more intuitive identifiers (e.g., "House #23"), reducing confusion and enhancing navigation.

To further involve users, the final section of the dashboard includes a feedback module. Here, users can indicate which explanation they found easier to understand and more trustworthy. They are also invited to leave open comments, enabling the design team to gather continuous input for iterative refinement.



💬 User Feedback

1️⃣ Which explanation do you find easier to understand?

● SHAP
○ LIME
○ Both equally
○ Neither

2️⃣ Which explanation do you trust more?

● SHAP
○ LIME
○ Both equally
○ Neither

🗣 Any additional comments or suggestions?

Figure 6: Feedback Module

Taken together, the final prototype successfully embodies the original design challenge: enabling users—especially those with a foundational understanding of data science—to critically assess and interpret AI-driven predictions for housing prices. The dashboard transforms abstract XAI theory into a usable, transparent, and reflective tool that supports trust and comprehension in high-stakes decision-making contexts.

7.3 Visualization of the Interaction of the User with the Concept in the Use-Context

The interaction design was grounded in the design thinking framework, emphasizing empathy with users, iterative prototyping, and continuous feedback. The user journey reflects a real-world scenario in which a data science student is attempting to understand an AI model's prediction for a specific house. Motivated by a desire to make an informed housing decision, the student selects a house, reviews its original features, and examines both SHAP and LIME explanations.

They analyze how the features contributed to the predicted price, using the waterfall and summary plots for SHAP, and LIME's localized feature impact visualization. Through the feature comparison table, the student observes where SHAP and LIME agree or differ and reflects on which explanation feels more trustworthy. If confusion arises, they consult integrated tooltips or the feature overview panel. After their exploration, they provide feedback, which contributes to refining the tool in future iterations.

This use-context illustrates how the dashboard supports informed decision-making by making complex AI predictions more understandable and transparent. It bridges the gap between model complexity and user comprehension, particularly for first-time homebuyers with a technical background, offering a robust platform for interpretability in high-stakes scenarios.

8 Short description of design archive

- *Method overview, references to the archive that contains materials used in user research (e.g. probe materials, interview guide, observation scheme), the notes you took throughout your design process and collected RAW data.*

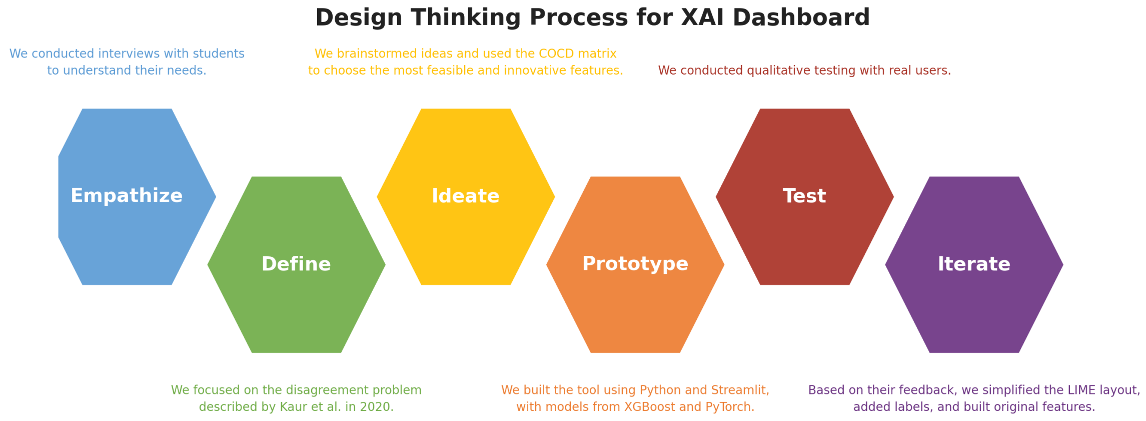

Design Thinking Process for XAI Dashboard
We conducted interviews with students to understand their needs.
We brainstormed ideas and used the COCD matrix to choose the most feasible and innovative features.
We conducted qualitative testing with real users.
Empathize
Define
Ideate
Prototype
Test
Iterate
We focused on the disagreement problem described by Kaur et al. in 2020.
We built the tool using Python and Streamlit, with models from XGBoost and PyTorch.
Based on their feedback, we simplified the LIME layout, added labels, and built original features.

Figure 7: Design Archive

The design archive serves as a comprehensive record of the development process behind the explainable AI dashboard prototype. It captures the progression from initial problem framing to ideation, user testing, and final implementation. At its core, the archive reflects a design thinking approach, evidenced by documented brainstorming sessions, early concept sketches, and prioritization of ideas using the COCD matrix method. These visual artifacts reveal how the team explored a wide solution space before converging on the most feasible and user-relevant features.

Supporting this creative process, the archive includes feedback gathered from qualitative user testing with data science students—specifically those representing the target audience of the dashboard. These materials include structured feedback tables summarizing participants' preferences, confusions, and improvement suggestions, as well as screenshots of the prototype in use during testing. This data was critical in shaping the user experience, especially around explanation clarity and interface language.

The archive also contains evidence of multiple prototype iterations, highlighting technical development through versioned Python scripts and Streamlit implementation files. Each iteration incorporated changes based on user input, such as improved tooltips, better labeling of house samples, and the removal of less interpretable explanation methods tied to neural networks.

Altogether, the archive illustrates a transparent and traceable design process grounded in iterative refinement, user feedback, and interdisciplinary collaboration. It provides a robust foundation not only for replicability but also for continued improvement of explainable AI systems aimed at non-expert users.

References

(CBS), Statistics Netherlands. 2024. "Housing Market Reports." CBS Netherlands. https://www.cbs.nl/en-gb.

Kaur, Harmanpreet, Harsha Nori, Simone Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. "Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools." In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14. ACM. https://doi.org/10.1145/3313831.3376219.

Research, Rabobank. 2024. "Housing Market Analyses Netherlands." Rabobank Economics. https://economics.rabobank.com/.

Statistiek, Centraal Bureau voor de. 2024. "Woningmarkt Dashboard." CBS.