Design Report

Interactive and Explainable AI

Stijn But Minji Kim Xuechun Lyu Sercan Şeref

2025-05-28

This report summarizes the design, development, and evaluation of an XAI dashboard for data science students, especially first-time home buyers. The dashboard addresses the "disagreement problem" in XAI, where different explanation methods (SHAP, LIME, Integrated Gradients, SmoothGrad, GradientShap) may provide conflicting feature importances. Using a user-centered, iterative design thinking process—Empathize, Define, Ideate, Prototype, Test, and Iterate—the team engaged real users to inform each phase. The dashboard enables side-by-side comparison of explanation methods, intuitive visualizations, and integrated user feedback. Findings show the dashboard clarifies feature importance, supports user trust, and highlights areas for further improvement. The design archive documents all stages, ensuring transparency and replicability. This work offers practical insights for developing accessible, trustworthy XAI tools for real-world decision-making in complex domains like the housing market.

Table of contents

1 Introduction

The disagreement problem 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. 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

The development of our XAI dashboard followed a user-centered, iterative design process, closely aligned with the design thinking framework: Empathize, Define, Ideate, Prototype, Test, and Iterate.

- **Empathize:** We began by conducting interviews and exploratory sessions with data science students—our intended users—to understand their needs, expectations, and trust issues regarding explainability tools. This early user research surfaced key pain points and preferences, forming the foundation for our design direction.
- **Define:** Drawing on these insights, we focused our problem definition on the "disagreement problem" in XAI, as described by Kaur et al. (2020), where different explanation methods provide conflicting feature importances. This step clarified our goal: to support users in interpreting and reconciling divergent explanations, especially for first-time home buyers with a data science background.

- **Ideate:** The team engaged in creative brainstorming, using techniques such as the COCD matrix to generate and prioritize ideas. This structured ideation ensured that our feature set—such as side-by-side explanation comparison, intuitive visual encodings, and user feedback integration—balanced technical feasibility with user value.
- **Prototype:** We translated these ideas into functional prototypes, building the dashboard using Python and Streamlit, with models from XGBoost and PyTorch. Early versions allowed users to explore real housing price predictions and compare multiple explanation methods.
- **Test:** Qualitative user testing was conducted with real students, who interacted with the dashboard and provided structured feedback on usability, clarity, and trust. Their input was systematically captured and analyzed, highlighting both strengths and areas for improvement.
- **Iterate:** Based on this feedback, we refined the dashboard—simplifying the LIME layout, improving labels, and adding features such as original value toggles. Each iteration was documented, ensuring traceability and continuous improvement.

To communicate our process and outcomes, we created summary posters and a video walkthrough, supporting both engagement and replicability. Throughout, the design archive ensured that every decision was traceable, laying a foundation for accessible and trustworthy XAI tools.

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 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), 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. (Ribeiro, Singh, and Guestrin 2016)

Another important advancement is SHapley Additive exPlanations (SHAP). 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. (Lundberg and Lee 2017)

Integrated Gradients 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. (Sundararajan, Taly, and Yan 2017)

Another notable method is SmoothGrad 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.(Smilkov et al. 2017)

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, a comparative study was conducted applying SHAP and LIME to real estate valuation models. The 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. (Özçelik and Yildirim 2022)

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.

Another contribution to this field is the work, 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. (De Nadai et al. 2016)

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. (Rabobank 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. An effective communication of explanations requires careful consideration of the audience's domain knowledge, making simplicity, visual clarity, and contextual relevance crucial design principles.(Molnar 2020)

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 potential users, we conducted exploratory user research centered around a prototype dashboard designed to enhance explainability in AI-driven housing price predictions. Although the dashboard had not yet been deployed in a real-world environment, participants interacted with interactive mockups and guided walkthroughs that closely simulated the intended user experience. This approach allowed us to evaluate early design concepts and explanation strategies through qualitative methods, including pilot testing, semi-structured interviews, and observational feedback.

Our target users—data science students—were intentionally selected for their foundational knowledge of AI and predictive modeling. This choice allowed us to engage with participants capable of critically assessing the interpretability and usability of various XAI methods. While young adults in general represent a key demographic for first-time home buying, their typically limited exposure to data science would have necessitated a drastically simplified dashboard. Achieving such simplicity with advanced tools like SHAP, LIME, Integrated Gradients, SmoothGrad, and GradientShap—each involving inherently complex visualizations and concepts—would have compromised the dashboard's depth and functionality. By focusing on a technically literate audience, we were able to explore more sophisticated interactions and nuanced explanations across multiple techniques, all while avoiding the need to overly dilute core concepts.

To ensure that the dashboard aligned with real user needs, we carried out a series of pilot testing sessions and interviews with our target group. The goal was to explore the challenges users face in interpreting housing price predictions, evaluate the clarity and utility of different explanation methods, and collect actionable feedback on usability and feature design.

Participants consistently indicated that understanding how housing price predictions are generated is inherently difficult—even when provided with model outputs. While the dashboard helped increase transparency, most users found that interpreting the explanation methods required effort, particularly during initial use. Nevertheless, the side-by-side comparison of explanation techniques emerged as a major strength. This feature made it easy to identify where SHAP and LIME aligned or differed, which helped users recognize and reflect on the "disagreement problem." This capability added depth to their understanding and built greater confidence in interpreting model behavior.

Visual clarity emerged as a critical component of user understanding. Participants emphasized the importance of color-coded graphs, clean layouts, and clear labels. Suggestions for improvement included adding more detailed legends, improving axis labeling, and simplifying LIME explanations. The top-5 feature comparison module was repeatedly cited as one of the most helpful elements of the dashboard. It allowed users to quickly see which factors most strongly impacted predictions and how these varied across explanation methods.

At the same time, several areas for refinement were identified. Users expressed difficulty understanding certain LIME outputs, particularly the representation of prediction values and technical jargon. There were repeated calls for clearer organization, more intuitive labeling, and simpler language throughout. Additionally, users proposed new features to improve accessibility and interactivity—such as descriptive titles, informative tooltips, and even a "build my own house" function to explore hypothetical predictions.

While the dashboard was initially perceived as most suitable for data science–savvy users, participants agreed that with modest improvements—like onboarding tips or embedded explanations—it could be made accessible to a wider, less technical audience.

These insights directly influenced our design decisions moving forward. Based on user preferences, we refined SHAP visualizations, enhanced the clarity of color legends, and reorganized the side-by-side comparison interface.

In terms of validation, participants unanimously agreed that the dashboard addressed a meaningful need: helping users make sense of AI-driven predictions in high-stakes domains like housing. Many stated they would consider using such a tool in real decision-making contexts. However, trust in the model itself remained nuanced. While transparency improved user understanding, some expressed concern that disagreement between explanation methods could erode their confidence in the predictions—suggesting that explainability, while necessary, may not always be sufficient to build trust.

In summary, pilot testing validated the dashboard's core strengths—particularly its visual clarity and comparative approach—while identifying clear opportunities for refinement. These insights guided our iterative development process and reinforced the importance of maintaining a user-centered approach throughout the design of explainable AI systems.

3 Define

As we moved into the ideation phase, we brought together insights from our technical exploration of explainability techniques and the qualitative feedback gathered during early user engagement. This synthesis allowed us to clearly define the problem space and make informed choices about which XAI methods would best support interpretability in the context of housing price prediction.

Our interviews and pilot testing with data science students revealed two primary user needs. First, users wanted to understand the influence of input features at both a general and individual level. They were interested in seeing how features like property size or location affected predictions across the entire dataset, as well as how those same features contributed to the predicted price for a specific house. Second, many users found it challenging to interpret conflicting explanations from different XAI methods. This issue, closely related to what Kaur et al. (2020) describe as the "disagreement problem," often led to uncertainty and reduced trust in the model's outputs. Users needed support in making sense of these differing perspectives.

To address these needs, we carefully examined a range of contemporary XAI techniques, evaluating them on their interpretability scope (local versus global explanations), compatibility with different model types, cognitive accessibility, and the quality of their visualizations—especially within our Streamlit-based environment. Through this process, we identified three core methods as most relevant for our user group: SHAP, LIME, and gradient-based techniques.

SHAP (SHapley Additive exPlanations) was selected for its strong theoretical foundation and its ability to provide both global and local insights. Its consistent, additive explanations—especially when expressed in monetary values—were found to be particularly intuitive for users trying to make sense of housing price predictions. LIME (Local Interpretable Model-agnostic Explanations) was chosen to complement SHAP by offering localized surrogate explanations that highlight how small changes in input features affect individual predictions. LIME's model-agnostic nature also made it suitable for cross-model comparison, which aligned well with our exploratory goals. Gradient-based methods, such as Integrated Gradients and SmoothGrad, were included specifically for deep learning models. While more complex, these techniques offered valuable insights into neural network internals for more advanced users.

This analysis led us to design a dashboard with two main tracks: one for XGBoost predictions using SHAP and LIME, and another for neural network predictions featuring SHAP, LIME, and gradient-based methods. This structure allowed us to balance breadth and focus, ensuring the interface remained accessible while still supporting a range of model complexities.

Synthesizing the findings from user research and our technical review, we defined a clear focus for ideation: how might we help data science students build trust in and understanding of machine learning predictions by enabling accessible, meaningful comparisons between different explanation methods? Rather than seeking a single "best" explanation, we framed interpretability as a process of sensemaking—one that benefits from exposure to multiple methods, structured comparison, and intuitive visualization.

Our ideation efforts therefore concentrated on supporting contrast and reflection between selected explanation types, such as through side-by-side visualizations of SHAP and LIME outputs, overlap analysis between feature attributions, and interface toggles that allow users to adjust the choice of model. This approach was not arbitrary or purely technical; it was shaped by a careful balance of user needs, cognitive limitations, the strengths of XAI methods, and practical interface considerations. As a result, the ideation phase advanced with clarity and purpose, grounded in real user insights and aimed at building truly user-centered explainability tools.

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 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 emerged directly from the convergence phase and was shaped by both user needs and the COCD matrix prioritization. At its core, the dashboard enables users to compare multiple explanation methods side by side for each housing price prediction. This comparative approach directly addresses the "disagreement problem" by making differences and agreements between methods explicit, helping users build a more nuanced understanding of model behavior.

To make the explanations accessible and engaging, we implemented intuitive color schemes throughout the visualizations. For example, positive and negative feature impacts are clearly distinguished using green and red, allowing users to quickly interpret whether a feature increases or decreases the predicted price. Each explanation method is accompanied by a brief, plain-language description, ensuring that users understand the logic behind the visualizations without needing to consult external resources.

The dashboard also incorporates simple, user-friendly elements such as emojis to indicate agreement or disagreement between explanation methods. This playful touch helps reduce cognitive load and makes the interface more approachable, especially for users who may be new to XAI concepts.

By focusing on clear visual encodings, concise explanations, and the ability to compare multiple explanation methods, the solution supports both novice and advanced users in making sense of complex model outputs. The design is intentionally streamlined, avoiding unnecessary technical jargon and emphasizing clarity, interactivity, and trust-building throughout the user experience.

5 Prototype

The preprocessing workflow begins with a raw numeric dataset, where missing values are handled using median imputation (as done in the code). Outliers are removed from non-binary features using Z-score filtering (abs(z-score) < 3), ensuring more stable model behavior. After outlier filtering, we merge back the binary features, retaining only clean rows.

The data is then split into features (X) and target (y), followed by a train/validation/test split. Next, we normalize all features using StandardScaler to apply Z-score standardization. This scaled dataset feeds into the training of an XGBoost Regressor (with n_estimators=80), which is finally evaluated on the validation set using R^2 score and Mean Squared Error (MSE).

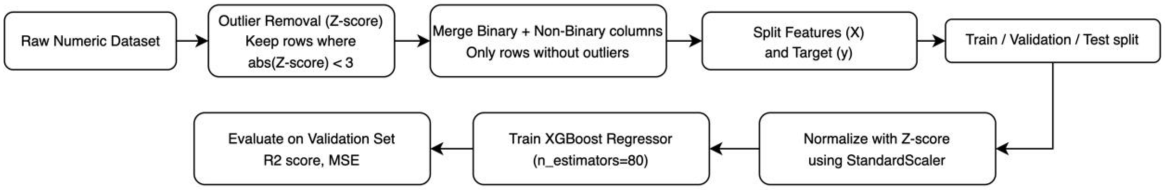
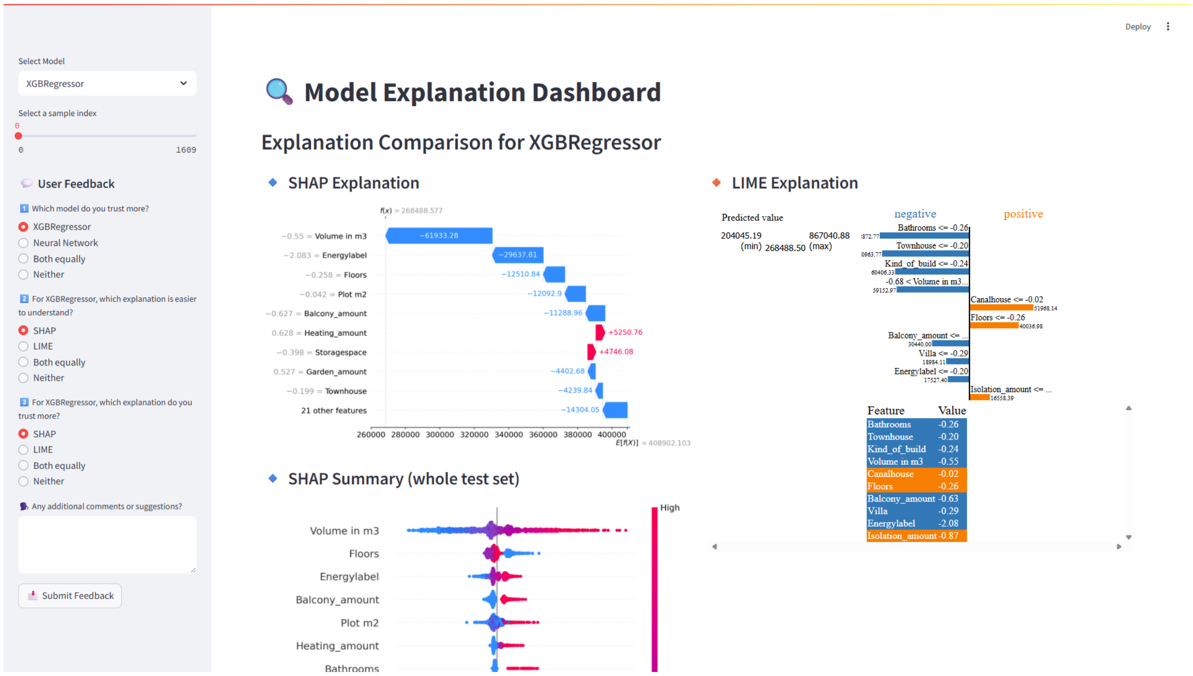

Raw Numeric Dataset
Outlier Removal (Z-score) Keep rows where abs(Z-score) < 3
Merge Binary + Non-Binary columns Only rows without outliers
Split Features (X) and Target (y)
Train / Validation / Test split
Normalize with Z-score using StandardScaler
Train XGBoost Regressor (n_estimators=80)
Evaluate on Validation Set R2 score, MSE

Figure 1: Preprocessing Workflow

To test our dashboard using Streamlit. This dashboard enables users to explore and compare the explanations generated by two powerful regression models—a tree-based XGBoost Regressor and a Neural Network—trained on a real-world housing price dataset. By integrating multiple state-of-the-art explanation techniques, the dashboard provides users with diverse and complementary insights into model decisions, fostering a clearer understanding of how predictions are formed.



Select Model

XGBRegressor

Select a sample index

0

0 1609

User Feedback

1 Which model do you trust more?

● XGBRegressor
○ Neural Network
○ Both equally
○ Neither

2 For XGBRegressor, which explanation is easier to understand?

● SHAP
○ LIME
○ Both equally
○ Neither

3 For XGBRegressor, which explanation do you trust more?

● SHAP
○ LIME
○ Both equally
○ Neither

Any additional comments or suggestions?

Submit Feedback

Model Explanation Dashboard

Explanation Comparison for XGBRegressor

SHAP Explanation

SHAP Summary (whole test set)

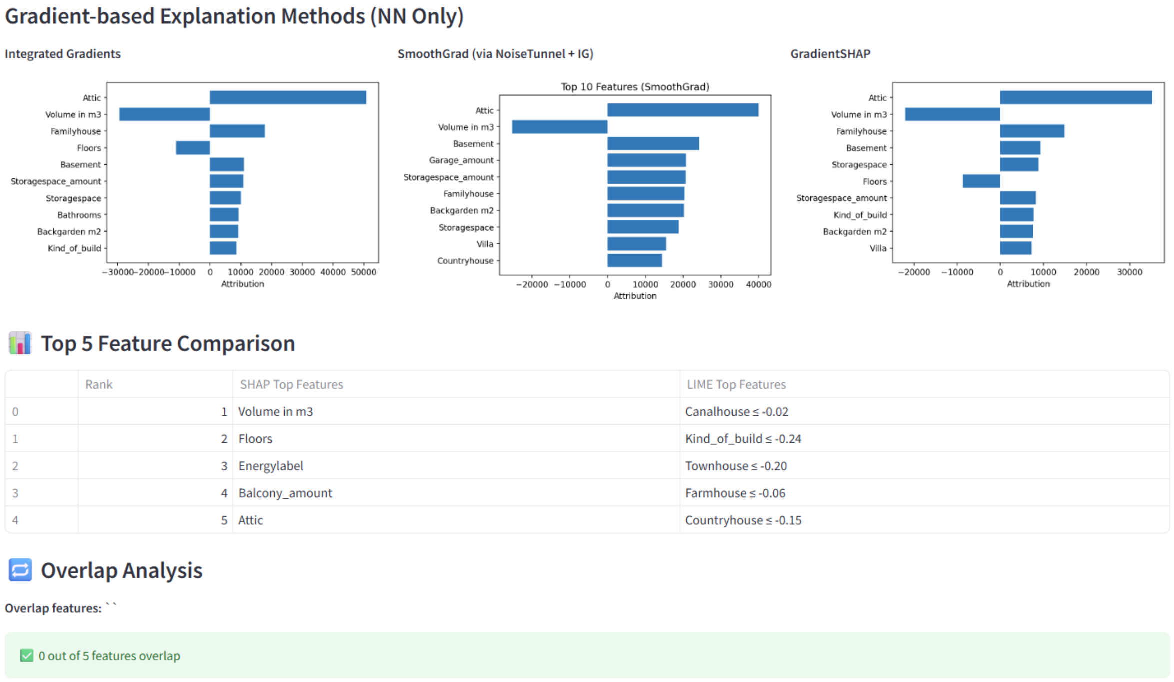
LIME Explanation

Feature	Value
Bathrooms	-0.26
Townhouse	-0.20
Kind_of_build	-0.24
Volume in m3	-0.55
Canalhouse	-0.02
Floors	-0.26
Balcony_amount	-0.63
Villa	-0.29
Energylabel	-2.08
Isolation_amount	-0.87

Figure 2: Protoype Dashboard Overview

Users begin their interaction by selecting the regression model they want to examine through a sidebar interface. Both models are trained on authentic housing data, and predictions are made on real test samples, ensuring that the explanations are grounded in practical, meaningful scenarios. Users can then select any test instance to analyze in detail, allowing them to investigate predictions on individual, realistic data points.

The dashboard incorporates three key explanation methods. SHAP (SHapley Additive exPlanations) offers both local and global perspectives: a local waterfall plot reveals the feature contributions for the chosen sample, while a global summary plot illustrates overall feature importance across the test set. LIME (Local Interpretable Model-agnostic Explanations) complements SHAP by generating interactive, HTML-based local explanations that approximate the model's behavior near the selected sample. For the Neural Network model, additional gradient-based methods—Integrated Gradients, SmoothGrad, and GradientSHAP—provide further interpretability by highlighting relevant features using advanced backpropagation techniques.



📊 Top 5 Feature Comparison

	Rank	SHAP Top Features	LIME Top Features
0	1	Volume in m3	Canalhouse ≤ -0.02
1	2	Floors	Kind_of_build ≤ -0.24
2	3	Energylabel	Townhouse ≤ -0.20
3	4	Balcony_amount	Farmhouse ≤ -0.06
4	5	Attic	Countryhouse ≤ -0.15

🔄 Overlap Analysis

Overlap features: ``

✅ 0 out of 5 features overlap

Figure 3: Explanation Methods for NN Tab

To enhance interpretability, the dashboard highlights the top five features identified by SHAP and LIME for the selected instance, and explicitly shows the overlap between these sets. This comparison helps users evaluate the consistency between different explanation methods and builds trust in the model outputs.

The user interface is designed to be intuitive and informative. Side-by-side visualizations display SHAP, LIME, and gradient-based explanation plots, enabling users to easily compare the strengths and nuances of each method. The sidebar also includes a feedback form where users can indicate which model and explanation they find most trustworthy and understandable, along with a space for open comments. This feature encourages engagement and provides insights into user preferences and comprehension, though in this final prototype phase it primarily serves as a record of user impressions rather than a mechanism for iterative development.

Overall, this prototype serves as a functional test environment aimed at learning from user interaction. It captures the essential steps of the user journey—model selection, explanation inspection, and trust feedback—and is designed to explore how different explanation techniques are received by users in realistic scenarios. By observing these interactions, we can better understand how to refine the dashboard to align more closely with user expectations and support effective decision-making in future versions.

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, 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 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.

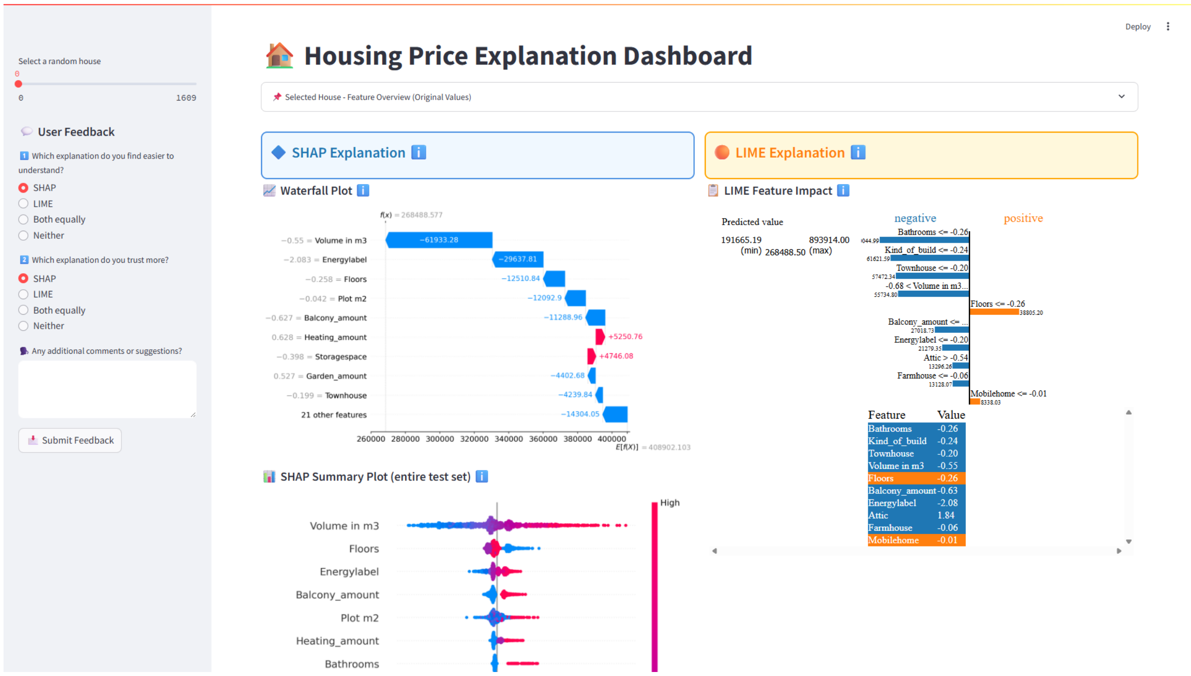
Additionally, we revised several tooltips to improve clarity and approachability, aligning them better with the data science students' level of familiarity. We also refined interface terminology—for instance, updating labels like “Select a random index” to more understandable phrases like “Select a random 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 Final Prototype

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.



1️⃣ Which explanation do you find easier to understand?

● SHAP
○ LIME
○ Both equally
○ Neither

2️⃣ Which explanation do you trust more?

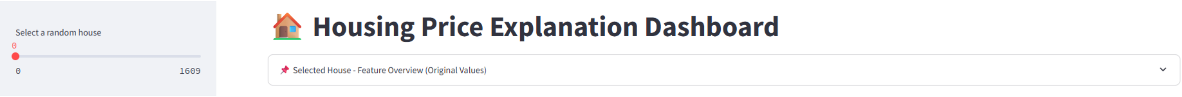
● SHAP
○ LIME
○ Both equally
○ Neither

Any additional comments or suggestions?

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: Dashboard Overview

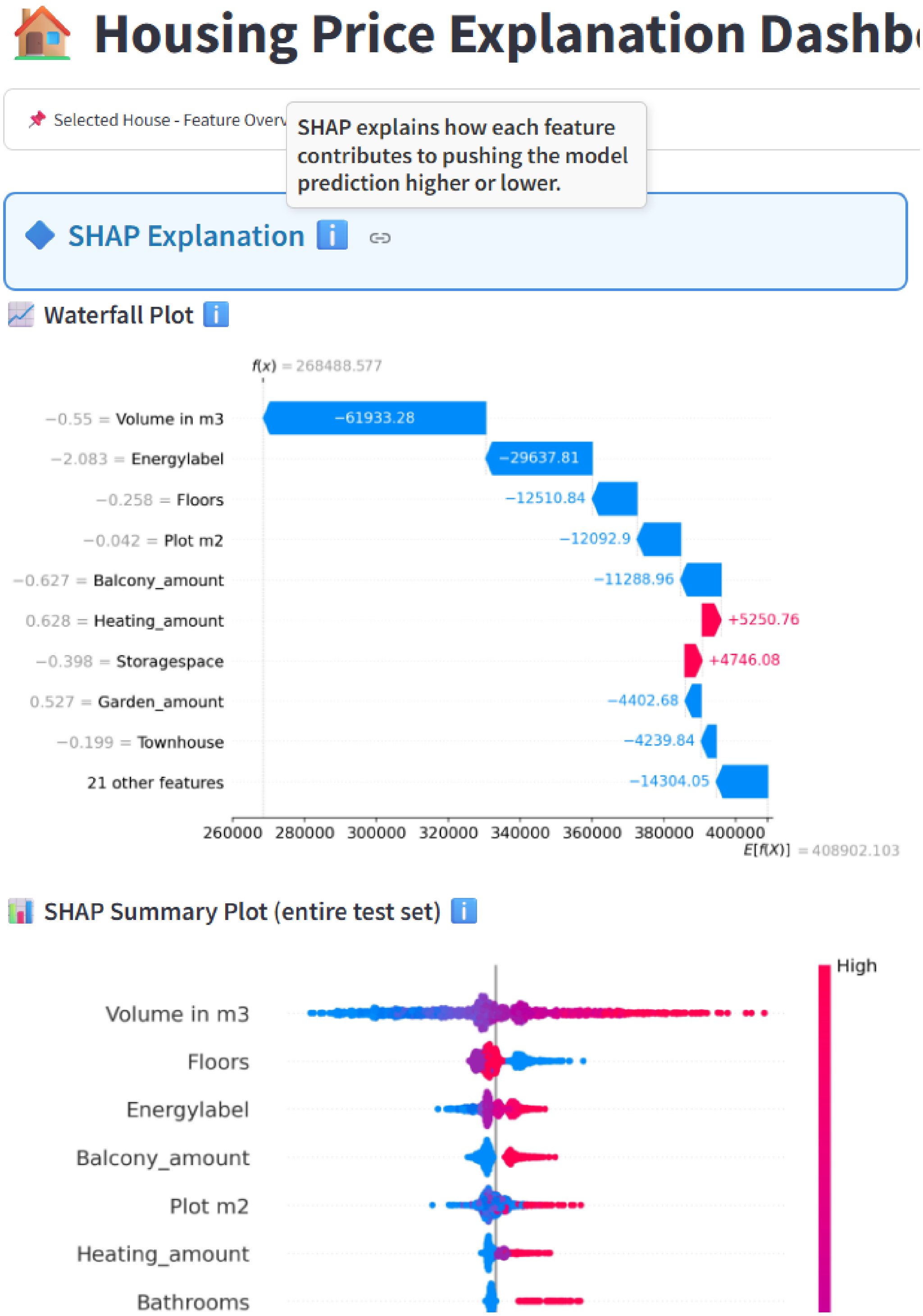
Upon launching the dashboard, users are welcomed by a streamlined interface that begins with the selection of a random house from the dataset. This selection dynamically updates the dashboard, triggering the rendering of explanations tailored to the chosen property. An expandable panel provides the original, 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 5: 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.

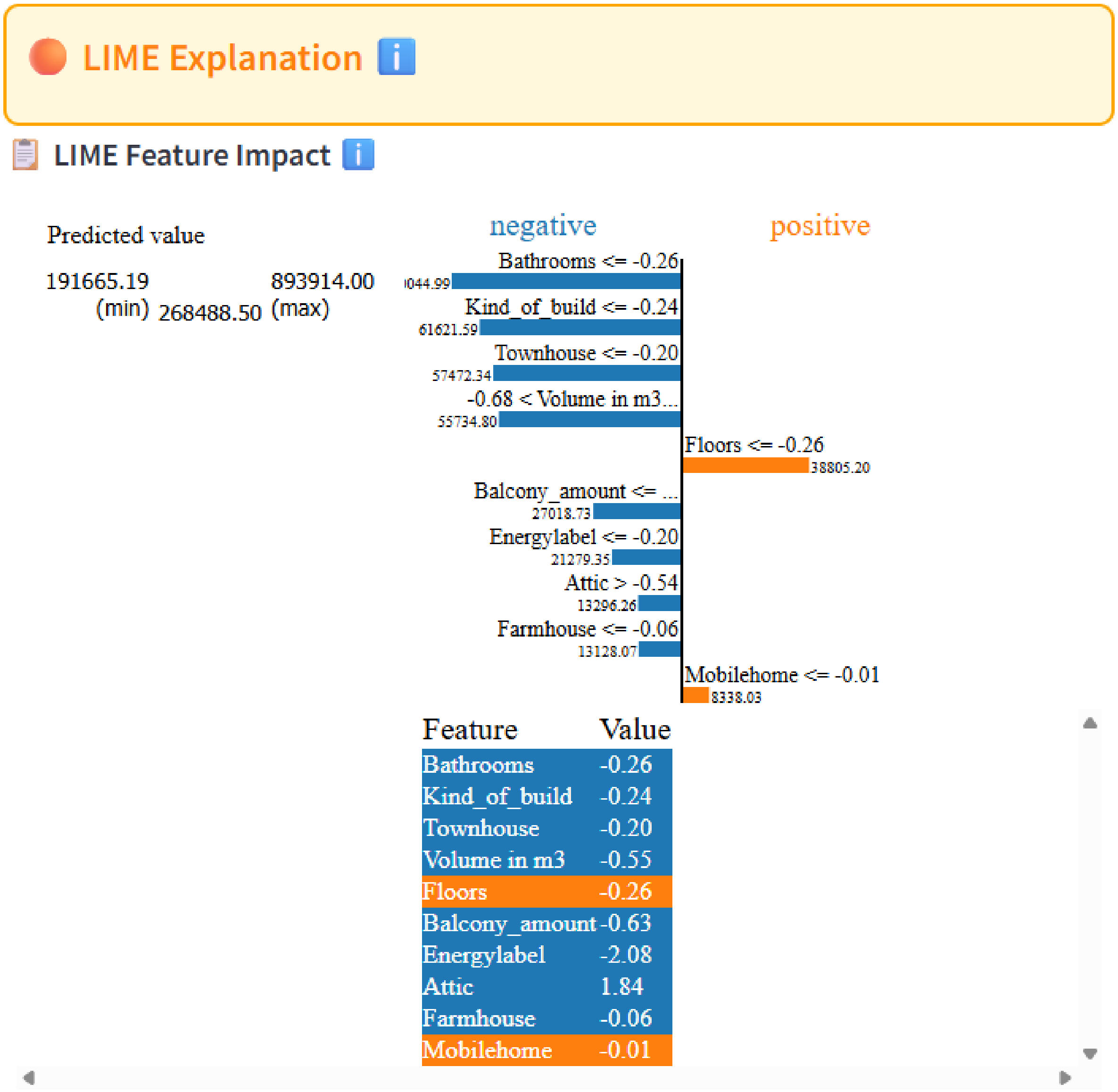
🔷 SHAP Explanation ℹ

📈 Waterfall Plot ℹ

📊 SHAP Summary Plot (entire test set) ℹ

Figure 6: 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.



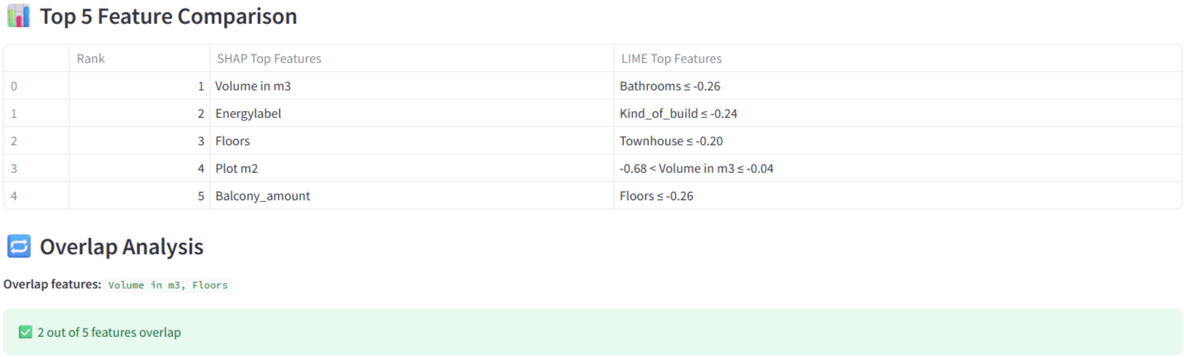
LIME Explanation

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 7: 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

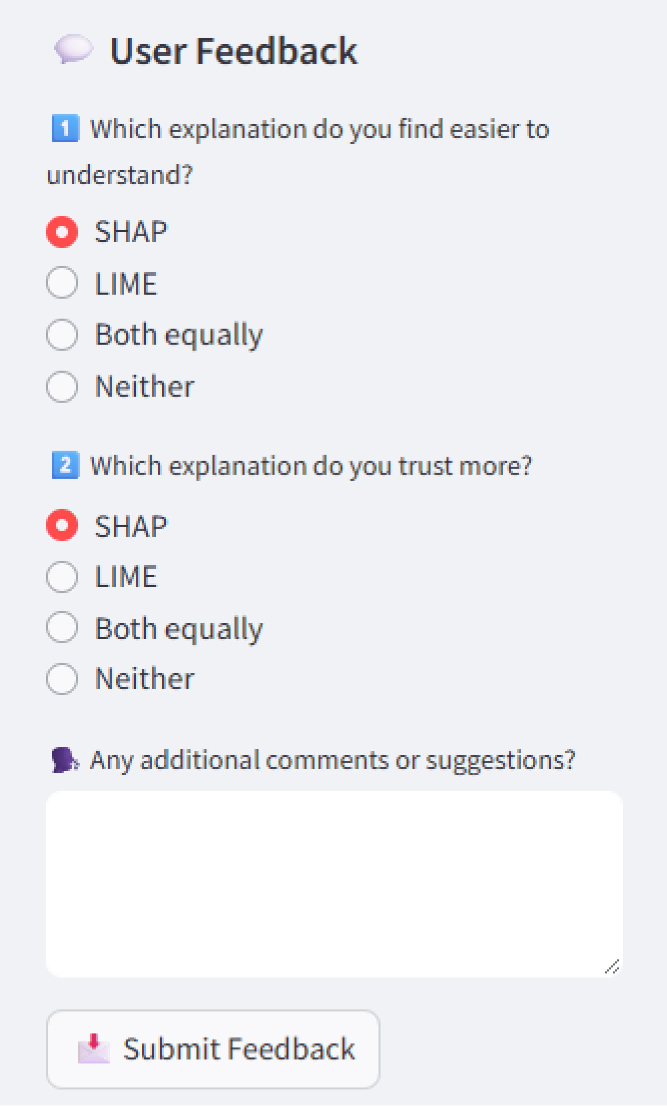
Overlap features: `Volume in m3, Floors`

✅ 2 out of 5 features overlap

Figure 8: 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, 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 9: 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 Interaction of the User

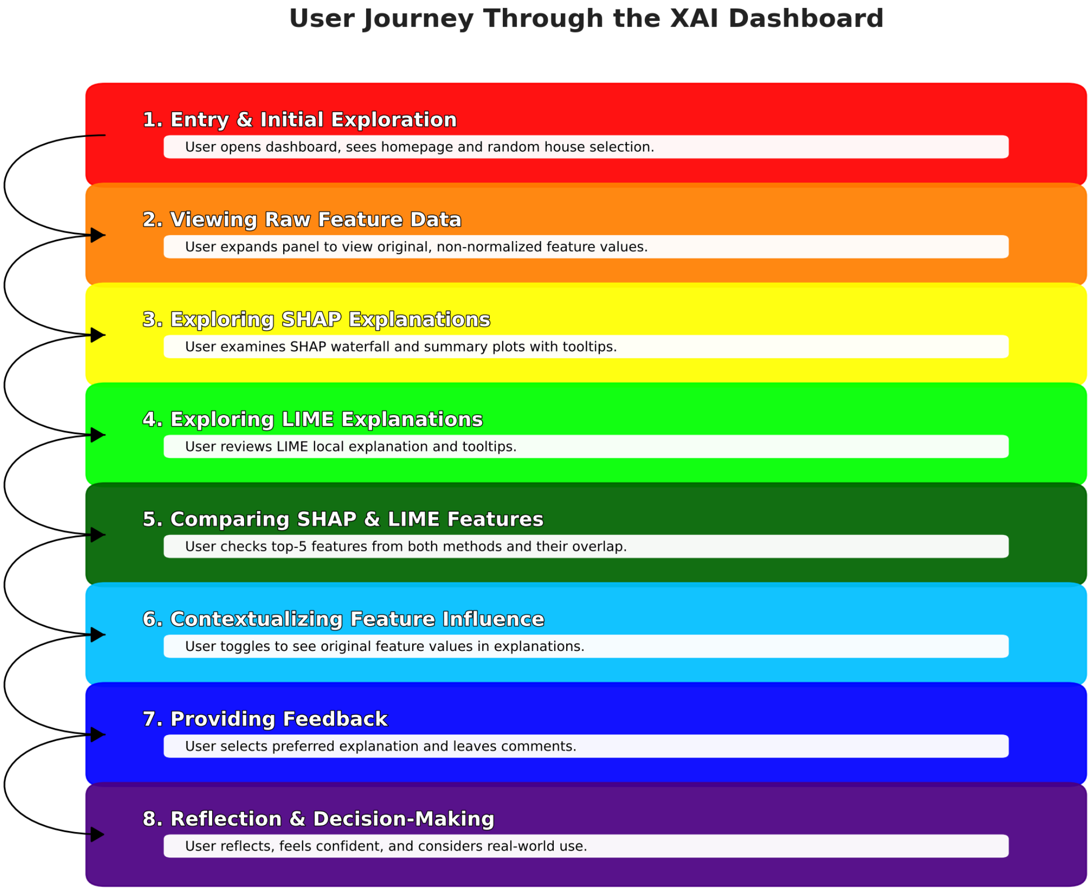

User Journey Through the XAI Dashboard
1. Entry & Initial Exploration
User opens dashboard, sees homepage and random house selection.
2. Viewing Raw Feature Data
User expands panel to view original, non-normalized feature values.
3. Exploring SHAP Explanations
User examines SHAP waterfall and summary plots with tooltips.
4. Exploring LIME Explanations
User reviews LIME local explanation and tooltips.
5. Comparing SHAP & LIME Features
User checks top-5 features from both methods and their overlap.
6. Contextualizing Feature Influence
User toggles to see original feature values in explanations.
7. Providing Feedback
User selects preferred explanation and leaves comments.
8. Reflection & Decision-Making
User reflects, feels confident, and considers real-world use.

Figure 10: User Journey

1. **Entry and Initial Exploration:** The user opens the dashboard and is greeted with a clean, intuitive homepage. They notice the option to select a house sample—by default, a random house from the dataset is already selected. This immediately gives them something concrete to explore without needing to make a choice first, reducing initial friction.

2. **Viewing Raw Feature Data:** Curious about the context, the user clicks to expand the panel showing the original, non-normalized feature values for the selected house. Seeing the raw data (e.g., living area in square meters, number of rooms) helps them ground the AI explanations in familiar, tangible terms.

3. **Exploring SHAP Explanations:** On the left panel, the user focuses on the SHAP waterfall plot. They observe how different features push the predicted price higher or lower. They hover over various plot elements and see tooltips explaining each feature's impact clearly in monetary terms. This visual helps them understand which features the model considered most influential for this specific house. They then glance at the SHAP summary plot below, which shows global feature importance across all houses. The color coding (red for high values, blue for low) immediately signals the general trends, reinforcing the local explanation they just saw.

4. **Exploring LIME Explanations:** The user shifts attention to the right panel, where LIME offers a complementary local explanation. They review the HTML-rendered visualization showing how slight variations in feature values influence the prediction. Tooltips assist them in grasping LIME's logic—that it approximates the model locally through perturbed samples.

5. **Comparing SHAP and LIME Features:** To deepen their understanding, the user examines the central feature comparison table. Here, the top five features from both SHAP and LIME are listed side-by-side, with overlapping features clearly highlighted. This comparison invites them to reflect on where the explanation methods align and where they differ, fostering critical thinking about model interpretability.

6. **Contextualizing Feature Influence:**The user toggles the option to reveal original feature values directly within the explanation visuals. Seeing "Living space: 140 m^2" next to its impact strengthens their ability to connect abstract explanation metrics with concrete property details.

7. **Providing Feedback:** After interacting with the explanations, the user reaches the feedback section. They select which explanation method they found easier to understand and more trustworthy, and leave a comment about their experience. This interaction not only contributes to iterative improvement but also reinforces the user's engagement with the tool.

8. **Reflection and Decision-Making:** Armed with this deeper, comparative understanding of how the models interpret housing features, the user feels more confident in the AI predictions. They appreciate the transparency and educational value of the dashboard and consider how such insights could be applied in real-world housing decisions or further data science projects.

8 Short description of design archive

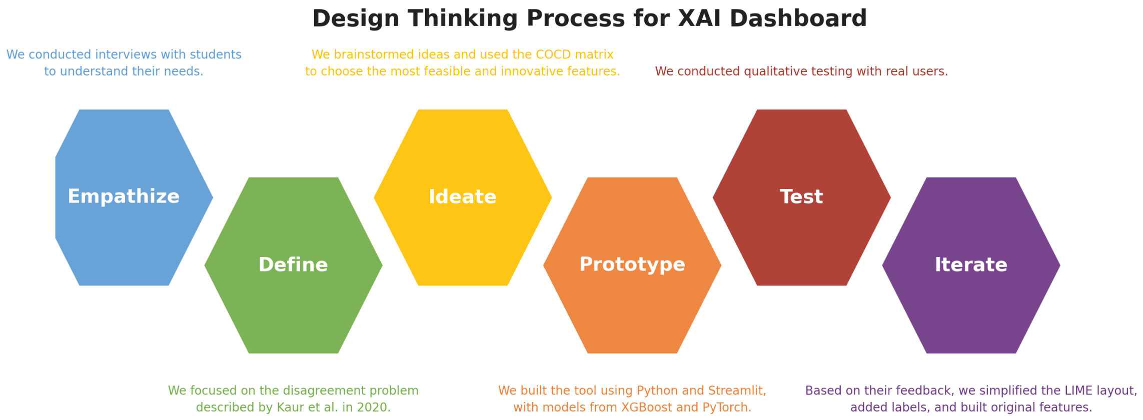

Design Thinking Process for XAI Dashboard
We conducted interviews with students to understand their needs.
We brainstormed ideas and used the COCD matrix to choose the most feasible and innovative features.
We conducted qualitative testing with real users.
Empathize
Define
Ideate
Prototype
Test
Iterate
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 11: Design Archive

The Design Archive documents the end-to-end development process of the interactive explainable AI (XAI) dashboard, following a user-centered, iterative design approach. It reflects how the team engaged with real users, explored creative solutions, developed functional prototypes, and refined them based on direct feedback—all while ensuring transparency, traceability, and technical soundness.

The process began with early pilot testing involving data science students, the intended users of the dashboard. These exploratory sessions were essential in surfacing initial expectations, trust issues, and preferences surrounding explainability tools. Feedback gathered during these conversations shaped the direction of ideation and feature prioritization. Detailed observations are captured in *Pilot Testing - Raw Transcript.docx*, offering a record of user reactions and reflections.

Building on these insights, the team entered the ideation phase, documented in *Creative Thinking - Assignment.docx*. This file contains brainstorming notes, initial sketches, and the application of the COCD matrix to evaluate and rank ideas. The document highlights the creative process used to balance technical feasibility with user value, ultimately guiding the design of the dashboard's core features—such as model selection, multi-method explanations (SHAP, LIME, gradients), and user feedback integration.

The first prototype was developed to bring these concepts to life and enable meaningful user interaction. This version, implemented in *combined_xai_dashboard.py*, was the foundation for structured user testing. It featured real model predictions on housing price data and allowed users to explore different explanation techniques. Accompanying this version, *Prototype1.png* and *Prototype2.png* capture screenshots of the dashboard during testing, providing visual documentation of the interface and user flow.

The results of the prototype's initial deployment are systematically captured in *User Test Results.xlsx*, which summarizes user feedback across categories such as Likes, Criticisms, Questions, Ideas, and Comments. This structured feedback was instrumental in diagnosing usability challenges, assessing the clarity of explanations, and identifying opportunities for refinement.

Based on this input, the team entered the iteration phase, resulting in an improved and more polished version of the dashboard, implemented in *Dashboard_FinalVersion.py.* This latest prototype integrated usability enhancements, improved visualizations, clearer interface language, and more intuitive interaction with explanation methods. A video walkthrough of this version, found in *XAI Dashboard.mp4*, demonstrates its final functionality and serves as a communication tool for both stakeholders and evaluators.

To support broader engagement and presentation, the archive also includes materials from the Design Market, where the project was pitched to peers and instructors. The file *Posters.docx* contains visual summaries of the project's goals, design rationale, process, and key findings, designed to clearly communicate the value and trajectory of the dashboard to a general audience.

In summary, the Design Archive provides a cohesive and comprehensive view of the development process, grounded in authentic user feedback and iterative refinement. By combining raw testing data, creative ideation artifacts, multiple code versions, and presentation materials, it ensures transparency, replicability, and a strong foundation for future development of explainable AI systems tailored to non-expert users.

References

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

De Nadai, M., J. Staiano, R. Larcher, N. Sebe, D. Quercia, and B. Lepri. 2016. "The Death and Life of Great Italian Cities: A Mobile Phone Data Perspective." *arXiv Preprint*. https://arxiv.org/abs/1609.01845.

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.

Lundberg, S. M., and S.-I. Lee. 2017. "A Unified Approach to Interpreting Model Predictions." *arXiv Preprint*. https://arxiv.org/abs/1705.07874.

Molnar, C. 2020. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable (2nd Ed.)*. https://christophm.github.io/interpretable-ml-book/.

Özçelik, M. H., and S. Yildirim. 2022. "Explainable Artificial Intelligence Techniques in Real Estate Valuation: A Comparative Analysis." *Computers & Industrial Engineering* 174: 108039. https://doi.org/10.1016/j.cie.2022.108039.

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

Ribeiro, M. T., S. Singh, and C. Guestrin. 2016. ""Why Should i Trust You?": Explaining the Predictions of Any Classifier." *arXiv Preprint*. https://arxiv.org/abs/1602.04938.

Smilkov, D., N. Thorat, B. Kim, F. Viégas, and M. Wattenberg. 2017. "SmoothGrad: Removing Noise by Adding Noise." *arXiv Preprint*. https://arxiv.org/abs/1706.03825.

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

Sundararajan, M., A. Taly, and Q. Yan. 2017. "Axiomatic Attribution for Deep Networks." *arXiv Preprint*. https://arxiv.org/abs/1703.01365.