Final Report

Interactive and Explainable AI

Stijn But   Minji Kim   Xuechun Lyu   Sercan Şeref

2025-04-21

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

- *Description of the conducted assignment*

The disagreement problem, as highlighted by Kaur et al. (2020), emphasizes the challenge of interpreting conflicting feature importances provided by different explanation tools and methods. This issue is particularly relevant in the context of dashboarding and explanation tools, where users often struggle with varying explanation styles and visualizations.

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 (2024), 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. 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

- *A well-argued and detailed description of the conducted pilot testing and qualitative user research methods (if any), review of XAI tools/techniques or literature*

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.(Ribeiro, Singh, and Guestrin 2016)

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.(Smilkov et al. 2017)

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.(Sundararajan, Taly, and Yan 2017)

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.(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, 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. (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.

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. (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. 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) ((CBS) 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. (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, Pilot Testing, and Results of Qualitative Analysis

In addition to reviewing and testing various explanation methods, we conducted a focused user research study with data science students—our primary target audience—to evaluate the perceived value of the dashboard concept. The feedback gathered from these students strongly validated the need for such a tool. Participants consistently indicated that being able to compare different explanation methods side-by-side would significantly enhance their ability to understand and interpret model predictions. This input directly influenced the design and features of our prototype dashboard, ensuring it aligns with the needs and expectations of data science students who are considering buying their first house.

2.3.1 Summarize Key Findings from User Research

The user research highlighted several important insights. The main challenge faced by students was understanding how housing price predictions are generated and which features most influence these predictions. The dashboard's side-by-side comparison of explanation methods (such as SHAP and LIME) was seen as especially valuable, as it allowed users to easily observe both agreements and disagreements between methods. Participants also emphasized the usefulness

of the top-5 feature comparison and the importance of clear, visually intuitive explanations. Overall, the research confirmed that such a dashboard meets a real need for transparency and interpretability in housing price prediction models, particularly for data science students preparing to make significant financial decisions.

2.3.2 Discuss Patterns or Themes

A recurring theme was a preference for SHAP explanations, mainly due to their global perspective and visually intuitive plots. Visual clarity—such as color-coded graphs and clear labels—was repeatedly mentioned as crucial for comprehension. Some users noted confusion with certain LIME outputs and suggested that improved labeling and explanations would enhance usability. The side-by-side comparison feature was universally appreciated, as it directly addressed the "disagreement problem" and made the dashboard's purpose clear.

2.3.3 Relate Findings to the Dashboard Design

Feedback directly influenced the dashboard's design: the side-by-side comparison module was prioritized, and efforts were made to improve visual clarity through color coding and concise labeling. Suggestions for further improvements, such as adding more descriptive titles and labels, were noted for future iterations. The dashboard's layout and features were refined to better align with user needs for transparency and ease of interpretation.

2.3.4 Address Limitations

The research was limited by a small and relatively homogeneous sample, primarily consisting of data science students. This focus aligns with our project's target audience—data science students who are considering buying their first house. While this ensures that the dashboard is well-tailored to users with some technical background, it may affect the generalizability of the findings to broader audiences, such as first-time home buyers without data science expertise. Additionally, some users found the explanations initially difficult to interpret, indicating a learning curve that could be mitigated with additional guidance or tutorials. Future iterations should consider expanding user testing to include a more diverse group to ensure broader accessibility and relevance.

2.3.5 Conclude with Validation

Overall, the user research validated the need for the dashboard. Participants agreed that it addresses a real need for understanding housing price predictions and appreciated the transparency provided by feature importance explanations. While some skepticism remained regard-

ing trust in the AI model—especially when different methods highlighted different features—the dashboard was seen as a valuable tool for interpreting and comparing model outputs.

2.3.6 Link to Future Work

Findings suggest that future development should focus on enhancing visual clarity, adding user guidance, and possibly introducing interactive features such as a "build your own house" simulator. Broader user testing with more diverse groups is recommended to ensure accessibility and relevance beyond data science students. These steps will help refine the dashboard and maximize its impact for all users interested in explainable AI for housing price prediction.

3 Define

- *A well-argued description of the XAI techniques and user needs chosen to focus on in the ideation phase*

In the ideation phase, the XAI techniques selected were carefully chosen based on their potential to address the disagreement problem highlighted in the work of Kaur et al. (2020). This problem centers around the difficulty users face when interpreting conflicting feature importances produced by different explanation methods. To tackle this, the team focused on a combination of techniques known for their complementary strengths in offering transparent and interpretable explanations.

One of the primary techniques selected was SHAP (SHapley Additive exPlanations), a game-theoretic approach that ensures consistency and fairness by considering all possible combinations of features. SHAP provides both global and local explanations, making it a versatile tool for understanding overall model behavior as well as individual predictions. In addition to SHAP, LIME (Local Interpretable Model-agnostic Explanations) was chosen for its ability to explain individual predictions through an interpretable surrogate model. Its model-agnostic nature and localized explanations made it particularly valuable for interpreting specific outputs.

The team also selected several neural network-specific techniques, starting with Integrated Gradients, a method that computes feature attributions by integrating the gradients of the model's output with respect to the input along a path from a baseline to the actual input. Integrated Gradients is especially effective for deep learning models, offering a clear and theoretically grounded attribution of input features. Building on this, SmoothGrad was incorporated to enhance the interpretability of gradient-based methods by reducing noise through the averaging of gradients over multiple noisy samples of the input. This leads to smoother and more comprehensible explanations. Finally, GradientShap was included as a variant of Integrated Gradients that combines the baseline approach with random sampling, averaging the attributions across multiple baselines to improve robustness and reliability.

The user needs identified during the ideation phase were grounded in qualitative research conducted with data science students, which revealed several key priorities. First, users expressed a strong desire to be able to compare multiple explanation methods side-by-side, helping them to better interpret differences and similarities between explanations and to build greater trust in the AI models. Visual clarity emerged as another critical need, with users emphasizing the importance of intuitive, color-coded graphs and the provision of simplified visualizations for novices alongside more detailed plots for more advanced users.

Beyond visual presentation, users highlighted the importance of accessibility and transparency. They needed brief, easy-to-understand explanations of each method to support their learning and ensure effective engagement with the dashboard, even for those without deep technical expertise. Interactivity was also seen as essential. Features like a "What if?" scenario explorer, trust meters to indicate confidence levels and data quality, and engaging animations were identified as ways to make the dashboard more dynamic, user-friendly, and supportive of deeper exploration.

Lastly, the relevance of the dashboard to real-world problems was seen as a major factor in its perceived value. By focusing on housing price predictions—a topic highly pertinent to the Dutch housing market and the everyday concerns of young adults—the project aligned the technical goals of explainability with practical, real-life decision-making needs.

4 Ideation

4.1 Description of the creative techniques used for divergence and convergence

- *Description of the creative techniques used for divergence and convergence*

Throughout the project, creative thinking techniques were systematically applied during both the divergent and convergent phases to generate, explore, and refine ideas for the dashboard prototype.

In the divergence phase, the team embraced an open brainstorming methodology based on the "Yes, and…" principle. This approach encouraged participants to freely build upon each other's suggestions without immediate critique, fostering a highly creative atmosphere. As a result, a wide array of ideas was generated, ranging from practical enhancements to bold, innovative features. Ideas included integrating multiple explanation methods side-by-side, using color-coded graphs (e.g., red for negative and green for positive influences), implementing emojis and animations to make complex results more intuitive, and developing features like a "What if?" simulator and a trust meter. To support these creative efforts, team members also engaged in activities like wireframing dashboard layouts, sketching interaction flows, and using visual metaphors—such as likening integrated gradients to a "dimmer switch"—to make abstract AI explanations more relatable.

Following the broad idea generation, the convergence phase was structured through a COCD Box (Creativity, Originality, Complexity, and Difficulty) framework. Here, ideas were categorized into Blue (feasible and easy to implement), Red (innovative and easy to implement), Yellow (innovative but harder to implement), and Grey (expensive or complex). This process enabled the team to filter the many brainstormed ideas and focus on those that offered the greatest impact relative to effort. Priority was given to feasible and innovative solutions such as the use of intuitive color schemes, brief textual explanations of each method, comparison modules between explanation techniques, playful visual enhancements (e.g., emojis and simple animations), and personalization options like simplified vs. detailed graphs for different users. The ideas from the Red and Blue zones were especially emphasized for rapid development, ensuring that the final product would be both creative and achievable within the project constraints.

By combining free, expansive idea generation with structured selection and refinement, the team effectively balanced innovation and practicality, ensuring that the final dashboard would be both technically sound and genuinely user-centered.

4.2 Description of the chosen solution

- *Result of the divergence and convergence technique - A well-argued description of the chosen solution*

The application of divergence and convergence techniques directly resulted in the design and development of a dashboard that was both highly functional and distinctly user-friendly.

One of the key outcomes was the explanation comparison module, which allowed users to view and compare feature attributions across multiple explanation methods, including SHAP, LIME, Integrated Gradients, and SmoothGrad. This feature directly addressed the disagreement problem (Kaur et al., 2020) by making it easier for users to observe where explanation methods agreed and where they diverged. As a result, users gained a clearer, multi-dimensional understanding of the AI model's behavior rather than relying on a single method's interpretation.

The use of intuitive color coding helped users quickly grasp the meaning behind the feature importances. Meanwhile, features like the trust meter and concise method explanations addressed the user need for transparency and helped foster greater trust in the AI's outputs.

Furthermore, by enabling users to toggle between simplified and advanced visualization modes, the dashboard successfully catered to the varying expertise levels of data science students who are first-time home buyers. This adaptability ensured that users, regardless of their technical proficiency, could effectively engage with the dashboard and gain valuable insights into housing price predictions.

The final solution also remained highly relevant to real-world problems. Given the sharp rise in housing prices in the Netherlands (CBS, 2024), understanding the factors that drive property

values has become increasingly important. The dashboard responded to this societal need by offering a practical, easy-to-use tool for interpreting housing price predictions, empowering users to make more informed financial decisions.

In conclusion, the creative divergence ensured that a rich and varied pool of ideas was explored, while convergence ensured that the most impactful, user-centric, and feasible ideas were executed. The resulting dashboard is a testament to the power of structured creative thinking, offering a technically advanced yet highly intuitive solution to a real-world challenge in AI explainability.

5 Prototype

- *A well-argued description of the developed prototype, including images of the prototype*

The developed prototype, implemented in python script, is an interactive, research-informed explainable AI (XAI) dashboard. It is designed to support data science students who are first-time home buyers in understanding and comparing how different machine learning models and interpretability techniques explain housing price predictions. The structure of the dashboard is directly guided by the findings of our user research and the XAI principles laid out in the literature.

The dashboard is built with Streamlit, chosen for its ability to rapidly create web-based interactive applications. The visual interface is intuitive and accessible to users with varying technical expertise, allowing seamless navigation between models and explanation techniques. The dashboard integrates SHAP, LIME, and Captum libraries, providing users with multiple explanation methods that emphasize both transparency and flexibility. SHAP enables both global and local model explanations using a game-theoretic framework, LIME offers localized model-agnostic explanations, and Captum supports gradient-based interpretation methods for neural networks.

The script starts by importing a comprehensive suite of packages essential for machine learning and explainability. In addition to data manipulation, and plotting, the script incorporates both tree-based (XGBoost) and neural network (PyTorch) models. It also supports explanation techniques across these models, enhancing comparative interpretability. The use of os, stats, and caching functions ensures efficient computation and robust statistical processing, including outlier detection.

The core of the prototype lies in its dual-model architecture. A neural network is defined using PyTorch, consisting of four fully connected layers with ReLU activations and dropout layers for regularization. This model mirrors the input dimensionality of the dataset (30 features) and serves as a counterpart to the XGBoost regressor, which is trained directly on standardized training data. The inclusion of both models enables users to explore the strengths and limitations of distinct machine learning paradigms and their corresponding explanations.

The data pipeline involves loading and cleaning a housing dataset. Non-binary features undergo outlier removal using z-score thresholds to ensure robust model behavior. Binary and non-binary features are recombined, and the resulting dataset is split into training and test sets. Standardization is applied to improve model convergence and ensure fair feature comparisons across explanation methods.

Following model training, the dashboard loads or reuses a pre-trained neural network (if available), enhancing performance and reproducibility. It provides users the ability to select either model for explanation and choose individual samples from the test set to explore. This sample-specific analysis aligns with user-identified needs for local interpretability and contextual understanding of individual predictions.

The user interface is structured around two primary display columns: one for SHAP and the other for LIME. For any selected sample, SHAP explanations are visualized using a waterfall plot to show individual feature contributions, and a summary plot that reveals overall feature importance across the test set. These visuals leverage clear, color-coded designs that were favored in user feedback sessions for their intuitive layout and interpretability.

On the LIME side, the dashboard generates and embeds interactive HTML explanations for the selected prediction. These explanations detail how each feature influenced the model's decision, helping users trace the decision logic for individual predictions.

When the neural network is selected, the dashboard additionally provides gradient-based explanations using Captum. It supports Integrated Gradients, GradientShap, and SmoothGrad, with bar plots displaying the top 10 features contributing to the model's output. These methods offer deeper insights into neural model behavior, particularly valuable for users curious about how black-box models arrive at their predictions.

The Comparison section synthesizes the top five features identified by SHAP and LIME, displaying them side by side and calculating feature overlap. This quantitative comparison helps users evaluate method agreement—a concern frequently cited in the XAI literature as the "disagreement problem." By presenting agreement metrics and prediction values side by side, the dashboard promotes critical thinking and confidence in model interpretation.

Finally, the dashboard includes a feedback section in the sidebar. Users can express which model and explanation method they found more understandable and trustworthy, and they are invited to share open-ended feedback. This mechanism supports iterative refinement of the tool, grounded in participatory design principles.

Below are screenshots illustrating the dashboard's core components:

SHAP Waterfall and Summary Plots Displays individual and global feature contributions for the selected model.

LIME Explanation Output Interactive HTML explanation of how features impacted a specific prediction.

Captum-Based Gradient Explanations (Neural Network) Feature importance derived from Integrated Gradients, SmoothGrad, and GradientShap.

Top Feature Comparison Between SHAP and LIME Table showing method agreement and differing perspectives on feature importance.

Overall, the python script represents a thoughtful synthesis of explainability research, user feedback, and machine learning best practices. By integrating multiple models and explanation methods into one cohesive, interactive tool, the dashboard not only educates users about AI decision-making but also empowers them to compare, question, and trust the predictions. It transforms abstract XAI theory into a practical resource for informed housing decisions, especially for audiences navigating this complex domain for the first time.

6 Test

- *A well-argued and detailed description of the conducted qualitative user research methods used for testing the prototype.*
- *Research question*
- *Results of the data analysis*

6.1 Qualitative User Research Methods

To evaluate the prototype dashboard, we conducted qualitative user research with data science students—our primary target audience—using a combination of structured user testing sessions and semi-structured feedback collection. Participants, consisting of both first-year and second-year master's students in AI, Data Science, and Computer Science, interacted freely with the dashboard, exploring features such as side-by-side explanation comparisons (SHAP, LIME, gradient-based methods), model selection, and feature importance visualizations. After hands-on use, participants provided feedback via a structured table capturing their affiliation, explanation method preference, likes, criticisms, questions, improvement ideas, and general comments. This approach enabled us to gather both quantitative data (e.g., method preferences) and rich qualitative insights (e.g., usability, clarity, and suggestions). Open-ended questions focused on preferences, perceived clarity, difficulties, and suggestions for improvement, ensuring that the research addressed interpretability, usability, and overall effectiveness of the dashboard for users with moderate technical knowledge but varying familiarity with explainable AI (XAI) techniques.

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.

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.

The data analysis followed a thematic coding process. User responses were reviewed and categorized into themes: preferred explanation method, clarity and usefulness of visualizations, confusion points, and suggestions for improvement. SHAP emerged as the most preferred

explanation method, praised for its intuitive layout, meaningful feature attribution, and ease of interpretation. In contrast, LIME received mixed feedback—some users appreciated its localized predictions, while others found it confusing or overwhelming due to the density of graphs and unclear terminology.

A recurring issue was confusion regarding the lack of agreement between methods, particularly between SHAP and LIME, with users questioning which method to trust and why discrepancies occurred. There was a strong demand for onboarding support, such as brief textual descriptions, a user guide, or tooltips explaining the methods and their roles. Suggestions also included clearer labels, more interactive help elements, and personalization (e.g., labeling houses rather than using index numbers).

Usability feedback highlighted minor interface issues, such as slow loading times, visual clutter, and a desire for more polished layout designs, especially for neural network explanations.

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, and improving interaction design to enhance overall clarity and trust in AI predictions.

7 Conclusion and Recommedations

- *Conclusion of the user testing*
- *A well-argued description of the final prototype, including visualizations*
- *Visualization of the interaction of the user with the concept in the use-context.*

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—primarily data science students—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 with descriptive labels like "House A" or "House in Amsterdam." A few participants also noted that switching between models could cause noticeable performance lags.

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 is an interactive Streamlit dashboard designed to help users understand how an AI model predicts housing prices, with a particular focus on making complex explanation methods accessible and comparable. Its main aim is to offer transparency and interpretability by integrating two widely used explainable AI techniques: SHAP and LIME.
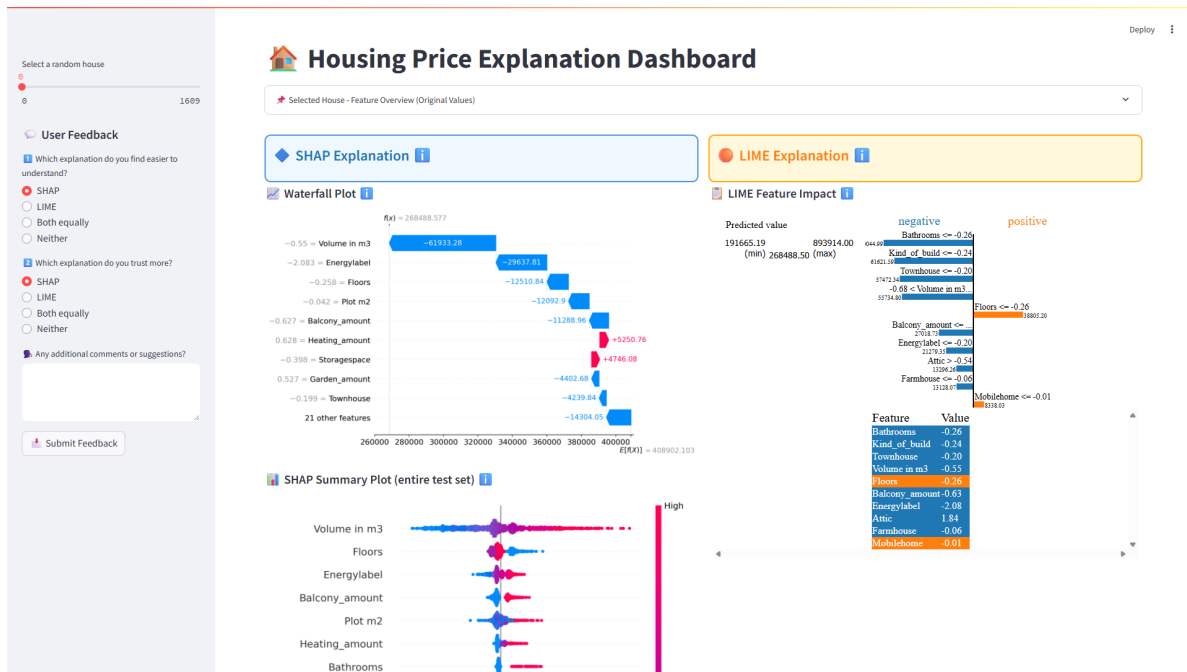
Figure 1: Dashboard Overview

When users launch the dashboard, they are greeted with a clean interface that begins by allowing them to select a random house from the dataset. This triggers the entire dashboard to update dynamically, centering all visualizations on the selected property. To provide context, an expandable panel shows the original feature values for that house, helping users ground the explanations in real data.
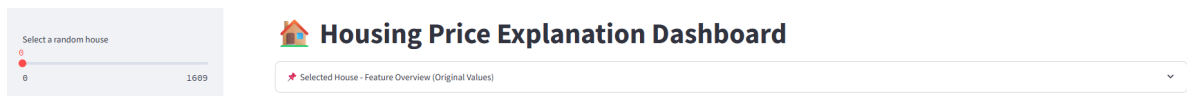

Figure 2: Random House Selection and Expendable Panel of Original Feautures

The explanation section is split into two visually distinct panels: SHAP on the left and LIME on the right. The SHAP section includes a waterfall plot that breaks down how each feature has influenced the prediction, showing whether it increased or decreased the estimated price. This is complemented by a summary plot, which shows global feature importance across all predictions, highlighting not just what mattered for one house but what matters generally across the dataset. Both visualizations are enhanced with hoverable tooltips that explain the visuals in simple, concise terms.
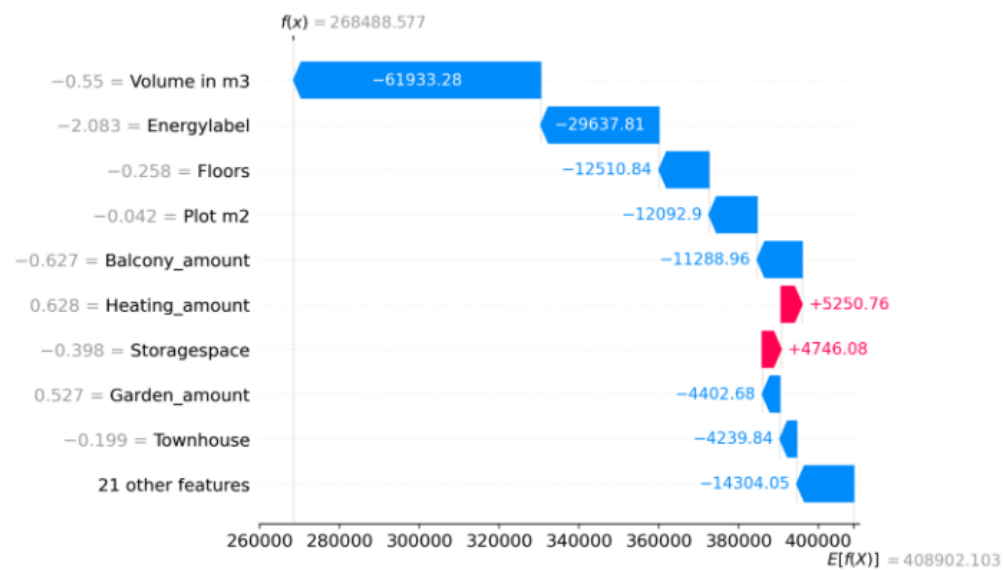
17

🏠 **Housing Price Explanation Dashb...**

📌 Selected House - Feature Overv...

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

◆ **SHAP Explanation** ℹ️ 🔗

📈 **Waterfall Plot** ℹ️

$f(x) = 268488.577$

$-0.55 =$ Volume in m3	-61933.28
$-2.083 =$ Energylabel	-29637.81
$-0.258 =$ Floors	-12510.84
$-0.042 =$ Plot m2	-12092.9
$-0.627 =$ Balcony_amount	-11288.96
$0.628 =$ Heating_amount	$+5250.76$
$-0.398 =$ Storagespace	$+4746.08$
$0.527 =$ Garden_amount	-4402.68
$-0.199 =$ Townhouse	-4239.84
21 other features	-14304.05

260000 280000 300000 320000 340000 360000 380000 400000

$E[f(X)] = 408902.103$

📊 **SHAP Summary Plot (entire test set)** ℹ️

Figure 3: SHAP Waterfall and Summary Plots

18

Opposite to SHAP, the LIME panel presents a local explanation specific to the selected house. Using an interactive HTML visualization, it highlights the features that had the most influence on that particular prediction, based on small input perturbations. As with SHAP, tooltips are integrated here to help users understand how the method works and how to interpret its output.
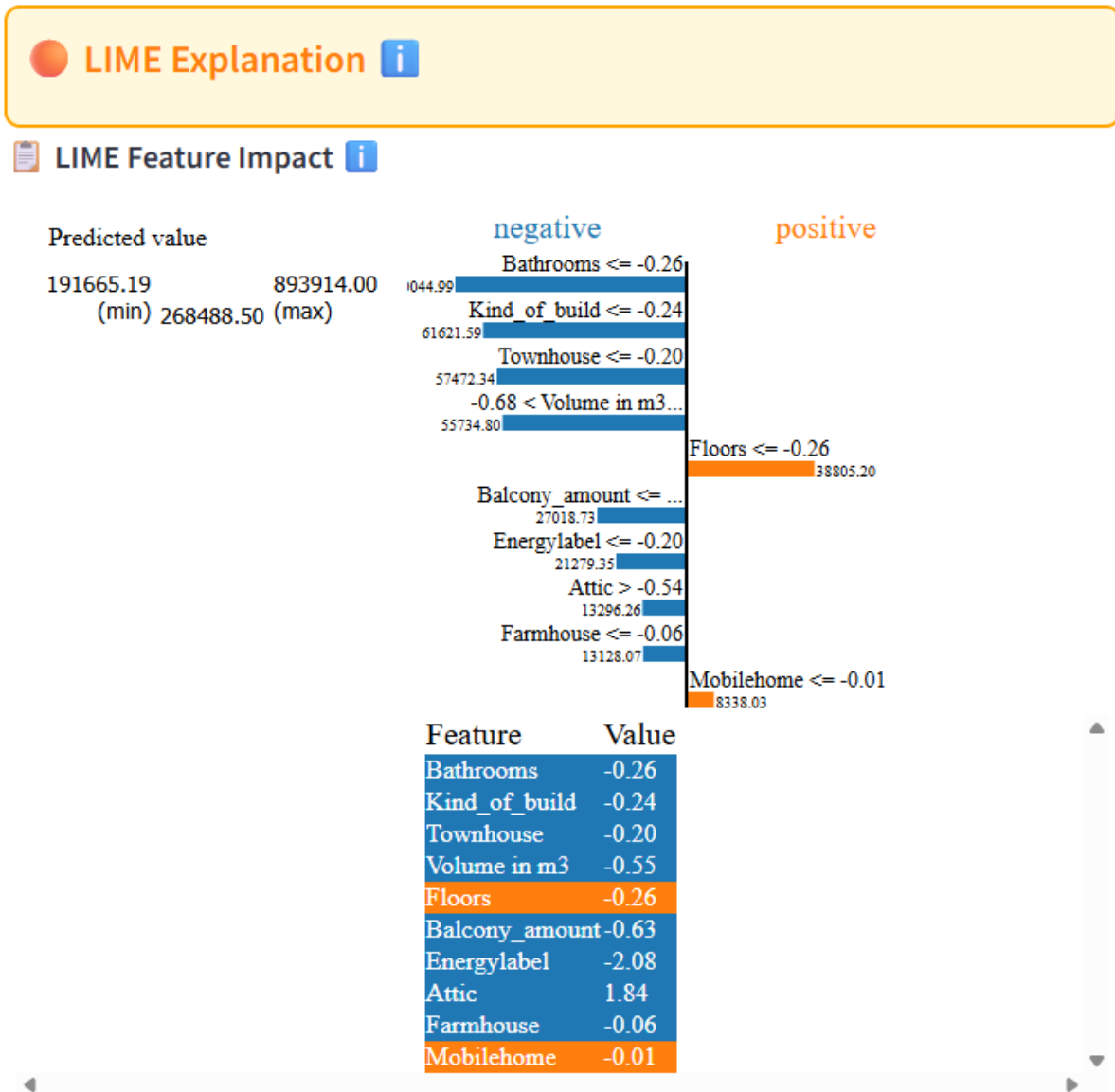

Figure 4: LIME Explanation Plots

To encourage critical thinking and enable reflection on the differences between methods, the dashboard includes a feature comparison table that displays the top five features identified by SHAP and LIME. Where these features overlap, the dashboard highlights this agreement; where they differ, users are prompted to consider why the explanations might diverge. This comparative design directly addresses the "disagreement problem" that often arises in XAI and was central to the project's original research question.

📊 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

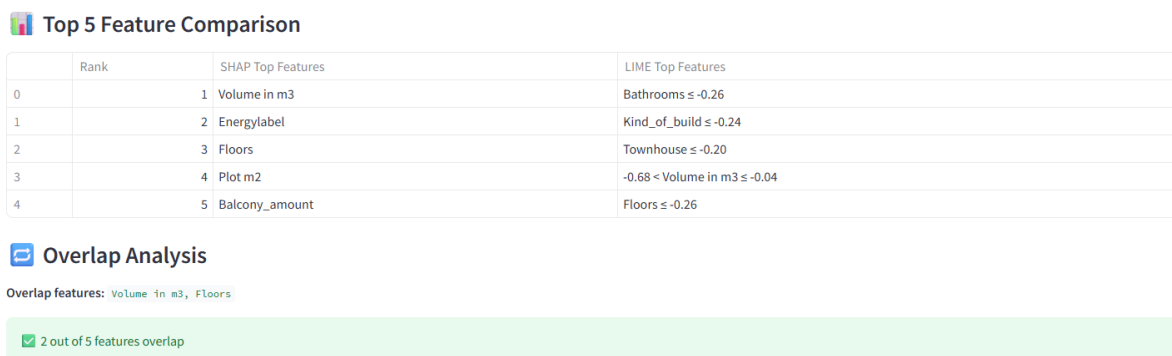**Overlap features:** `Volume in m3, Floors`

✅ 2 out of 5 features overlap

Figure 5: Top Feature Comparison Between SHAP and LIME

A key improvement made during development was the ability for users to view the original value of each feature. This added layer of context allows users not only to see how much a feature influenced the prediction, but also to understand which specific value caused that effect. Additionally, changes were made to improve overall usability, including replacing abstract index labels with clearer house identifiers and revising tooltip language to be more user-friendly.

User engagement is further encouraged through a feedback module embedded in the sidebar, where users can indicate which explanation method they found easier to understand and more trustworthy. They are also invited to leave comments, ensuring that the dashboard remains responsive to user needs and continues to improve over time.

Figure 6: Feedback Module

Overall, the final prototype strikes a balance between technical depth and accessibility. It empowers users—especially those with a foundational understanding of data science—to explore, compare, and contextualize AI predictions in an intuitive environment. The result is a practical and educational tool that supports better decision-making in the high-stakes context of housing.

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.*

9 References

- Centraal Bureau voor de Statistiek. (2024). Woningmarkt Dashboard. CBS. Retrieved April 26, 2025, from https://www.cbs.nl/nl-nl/visualisaties/dashboard-economie/woningmarkt
- De Nadai, M., Staiano, J., Larcher, R., Sebe, N., Quercia, D., & Lepri, B. (2016). The death and life of great Italian cities: A mobile phone data perspective. arXiv preprint arXiv:1609.01845. https://arxiv.org/abs/1609.01845
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. arXiv preprint arXiv:1705.07874. 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., & Yildirim, S. (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
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. arXiv preprint arXiv:1602.04938. https://arxiv.org/abs/1602.04938
- Smilkov, D., Thorat, N., Kim, B., Viégas, F., & Wattenberg, M. (2017). SmoothGrad: Removing noise by adding noise. arXiv preprint arXiv:1706.03825. https://arxiv.org/abs/1706.03825
- Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. arXiv preprint arXiv:1703.01365. https://arxiv.org/abs/1703.01365
- Statistics Netherlands (CBS). (2024). Housing market reports. CBS Netherlands. https://www.cbs.nl/en-gb
- Rabobank Research. (2024). Housing market analyses Netherlands. Rabobank Economics. https://economics.rabobank.com/

(CBS), Statistics Netherlands. 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.

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.

Research, 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.

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