Final Report

Interactive and Explainable AI

Stijn But Minji Kim Xuechun Lyu Sercan Şeref

2025-04-21

This is XAI Report

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. (**rabobank2023?**) ((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 and Pilot Testing

In addition to reviewing and testing these methods, we conducted a brief user research study with data science students to assess whether they perceive value in such a dashboard. The feedback from this research validated the idea, as students expressed that a tool enabling the comparison of different explanation methods would be highly beneficial for understanding and interpreting model predictions. This validation further reinforced the relevance of our project and guided the design of the prototype dashboard.

papers about the housing problem

- *Results of the (qualitative) analysis / conclusion.*

Summarize Key Findings from User Research:

Highlight the main insights from your user research with data science students. For example, mention how students found value in comparing explanation methods and how this aligns with their needs for understanding housing price predictions. Discuss Patterns or Themes:

Identify recurring themes or patterns in the feedback. For instance, students might have expressed a preference for specific explanation methods (e.g., SHAP or LIME) or emphasized the importance of visual clarity in the dashboard. Relate Findings to the Dashboard Design:

Explain how the feedback influenced your design decisions. For example, if students valued side-by-side comparisons, describe how this feature was incorporated into the prototype. Address Limitations:

Acknowledge any limitations in your research, such as a small sample size or limited diversity in the user group, and how these might affect the generalizability of your findings. Conclude with Validation:

State whether the research validated the need for the dashboard and how it supports the relevance of your project. For example, emphasize that the feedback confirmed the dashboard's potential to help users interpret housing price predictions effectively. Link to Future Work:

Briefly mention how the findings could guide further development or testing of the dashboard, such as refining specific features or conducting more extensive user testing.

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, while the integration of playful elements like emojis and animations made the

dashboard more engaging and emotionally accessible—especially for non-expert users. 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 different levels of user expertise, from novice first-time home buyers to more technically proficient data science students. This adaptability was crucial in ensuring that users felt neither overwhelmed nor underserved by the dashboard, regardless of their background.

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*

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*

7 Conclusion

- *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.*

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.

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.