Designing Resilient AI-based Robo-Advisors: A Prototype for Real Estate Appraisal

Max Schemmer, Patrick Hemmer, Niklas Kühl, and Sebastian Schäfer

Karlsruhe Institute of Technology, Karlsruhe, Germany
{max.schemmer, patrick.hemmer, niklas.kuehl}@kit.edu
sebastian.schaefer@student.kit.edu

Abstract. For most people, buying a home is a life-changing decision that involves financial obligations for many years into the future. Therefore, it is crucial to realistically assess the value of a property before making a purchase decision. Recent research has shown that artificial intelligence (AI) has the potential to predict property prices accurately. As a result, more and more AI-based robo-advisors offer real estate estimation advice. However, a recent scandal has shown that automated algorithms are not always reliable. Triggered by the Covid-19 pandemic, one of the largest robo-advisors (Zillow) bought houses overvalued, eventually resulting in the dismissal of 2,000 employees. This demonstrates the current weaknesses of AI-based algorithms in real estate appraisal and highlights the need for troubleshooting AI advice. Therefore, we propose to leverage techniques from the explainable AI (XAI) knowledge base to help humans question AI consultations. We derive design principles based on the literature and implement them in a configurable real estate valuation artifact. We then evaluate it in two focus groups to confirm the validity of our approach. We contribute to research and practice by deriving design knowledge in accordance with a unique artifact.

Keywords: Robo-Advisors · House Price Prediction · Resilience · Explainable Artificial Intelligence

1 Introduction

Buying a house is one of the major life events of a person [13]. One of the most important questions is whether a house is appropriately valued, i.e., the house is offered within a price range that reflects its assets and is comparable to similar houses in terms of characteristics of the property and its surroundings. Due to the manifold factors impacting a property's value, its correct appraisal is challenging [9]. Therefore, more and more companies provide AI-based support, so-called robo-advisors, in real estate appraisals. The latest business model extension is that robo-advisors use their own appraisal algorithms to buy and sell houses (called iBuyer). One of these iBuyers, Zillow, experienced an external shock in the Covid-19 pandemic due to over-reliance on their AI advice. The AI systematically overvalued houses offered for sale on the market resulting in

the automatic acquisition of many overpriced houses. This eventually resulted in such immense losses that the whole business unit had to be closed, accompanied by the dismissal of a quarter of the employees [7]. One conscious design decision of the real appraisal algorithm was that human intervention was not intended [1]. However, this decision has turned out to be a mistake in disruptive times, such as a pandemic, as no human experts adjusted the house price forecasts. Research has shown that human adjustments in the context of uncertainty and existing unique human contextual information can help make systems more resilient [2]. To conduct positive adjustments, humans need to judge the quality of AI advice. To enable experts to do so, they need information on the knowledge base and reasoning of the AI. To address this requirement, we draw from the research stream of explainable artificial intelligence (XAI) [5]. XAI aims to open the "black-box" of AI and provides the user with understandable insights into the AI's decision-making [5]. Based on XAI literature, we derive design knowledge and develop a configurable artifact for resilient robo-advisors in the context of real estate appraisal with the goal to better enable users to assess the quality of the suggested valuation. Subsequently, we evaluate our artifact in two focus groups that indicate the usefulness of our design. In future research, we additionally aim to evaluate its efficacy quantitatively.

2 Design of the Artifact

2.1 Design

We derive four design principles (DPs) according to Gregor et al. [8]. They are based on existing justificatory knowledge and are instantiated as design features (DFs) in the artifact. Research has shown that psychological factors can prevent the user from effectively adjusting AI predictions [4]. Especially, the anchoring effect and information overload are important to consider [4, 10]. The anchoring effect refers to a cognitive bias that describes how the user builds predictions based on reference points, e.g., the AI's advice [4]. If the anchor effect is too strong, the user may have difficulties adjusting the AI prediction. Information overload describes the phenomenon that the user is overwhelmed by the amount of information which reduces their processing capability [10]. Therefore, we formulate:

DP1: *Provide the system with a configurable interface that shows the information on demand to prevent anchoring and information overload.*

Research has shown that providing humans with information on the uncertainty of the AI helps to judge the quality of the AI advice [17]. Therefore, we formulate:

DP2: *Provide the system with a prediction uncertainty measurement to judge the quality of the AI advice.*

Sensitivity analyses offer the possibility to analyze how different values of input features influence the AI's prediction[16], e.g., changing the number of bedrooms, allowing to evaluate the robustness of the prediction iteratively. Therefore, we formulate:

DP3: *Provide the system with capabilities to perform a sensitivity analysis to explore the AI's solution space interactively.*

Another possibility to evaluate the quality of the AI's decision is to provide the user with insights into the AI's decision-making process, which allows them to compare it with their own reasoning. An analogy is an interaction between a human advisor and a client. To assess the advice's quality, the client will ask the advisor to express the reasoning behind a proposed decision. Based on the insights, the decision-maker determines whether to rely on the advice or not. We hypothesize that a similar logic should hold for a robo-advisor. Therefore, insights in the decision-making process can be seen as discrimination support for the decision-maker to assess whether to rely on the AI [3]. Therefore, we formulate:

DP4: *Provide the system with explanations for the AI's reasoning to enable appropriate reliance on AI advice.*

2.2 System Overview

Figure 1 displays the DPs and their instantiation through resulting DFs. We display the information about all relevant features of a property by default. The information is presented in the table's left part as name-value pairs and an image of the house in the right corner. To address DP1, the user can display AI predictions and explanations on demand. This design feature prevents the human from anchoring on predictions and explanations of the AI and aims to reduce the information load (**DF1**).

To provide the user with intuitive information on the reliability of a specific prediction (DP2), its confidence, displayed as a 95% confidence interval, is communicated alongside the prediction (**DF2**).

To address DP3, we implement a what-if analysis [11, 16] (**DF3**). The values of the given instance can be adjusted as displayed in Figure 1. The user can change numerical values by using a slider. This slider uses an interval ranging from the minimum to the maximum value of the respective feature occurring in the data set. The difference to the base instance is presented with blue and red color-coding. The two buttons located in the header can be used to refresh the prediction or to reset the values. The resulting interactive prediction is highlighted in a yellow box.

To provide the user with insights into the decision-making process, the following XAI method is used (DP4): We implement the feature importance algorithm LIME (**DF4**) [12]. It displays the degree and direction each feature contributes to a decision made by the AI. The direction can be either in favor or against the prediction. Color-coded bar graphs are used to visualize these values.

The second method provided in our prototype is an example-based explanation approach (**DF5**). Research has shown that especially example-based explanations are intuitive for humans and contribute to better understanding the AI's reasoning [6]. The idea is to offer the user the possibility to compare a given instance with reference examples. Working with real-world examples similar to the base instance can help humans better understand whether the AI's prediction is

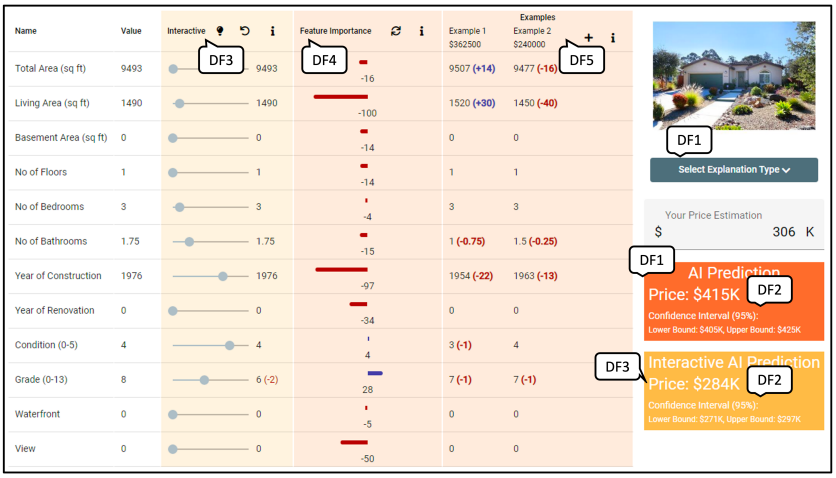

Fig. 1. The graphical user interface of the proposed prototype.

realistic. In addition, further examples provide insights into the knowledge base of the AI. The user can interactively select as many examples as desired.

Regarding the specific instantiation of the artifact, we choose a random forest regression as a model. The random forest has a mean average error of 104,000$ on a test set in which the house prices range until 7,700,000$. The instantiated artifact is accessible as an online application. The backend is based on the programming language Python with Flask as a framework. The data is stored in a MongoDB, while the frontend is based on Angular.

2.3 Demonstration

Figure 1 highlights the functionalities of the prototype with an example. In the example, the AI predicts a house price of 415,000k$. However, the historical market value of the home is below this value at 312,000$. Therefore, the user should be able to determine that the AI prediction has a low quality and should adjust it. The user in Figure 1 enabled all possible DFs—the prediction and uncertainty information, the sensitivity analysis, feature importance, and examples. Furthermore, the user performed a sensitivity analysis and slightly decreased some input factors. After these small changes, the AI predicts a considerably lower price of 284,000$. In addition, the feature importance shows a counterintuitive result. Both examples display cases very similar to the analyzed house but at a lower price (362,500$ and 240,000$). Overall, the sensitivity analysis, the feature importance, and both examples indicate an overvalued prediction of the AI and point towards an adjustment towards a lower house price. We argue that our design knowledge should help assess the quality of an AI prediction in real-world situations and enable users to adjust it accordingly.

3 Evaluation and Outlook

To evaluate the prototype, we used exploratory focus groups [15]. In general, focus groups are valuable for design research projects because they allow for direct interaction with respondents and for collecting large amounts of rich data [15]. In addition, exploratory focus groups aim to generate formative feedback for artifact refinement. We used a digital whiteboard to capture and structure the feedback. Our goal was to discuss the usefulness of the proposed DFs and thereby generate feedback for the refinement of the artifact. For the selection of the participants, we used purposeful sampling [14]. We searched for participants that have knowledge of human-AI collaboration and are at a representative age for buying a house. In total, 13 participants and two researchers participated in two sessions. We conducted two focus groups, one with a smaller group to facilitate more extensive exchange and one with a focus on collecting broad ideas. For both focus groups, the moderator was one of the primary researchers. Each session lasted 90 minutes. We first demonstrated the prototype to the participants during each session and led them through the different options using a click-through approach. Then, we let them freely use the prototype on their own. Next, we iteratively discussed three prediction examples, each with varying prediction quality. After each property, we collected feedback in a structured way by asking the participants to note down how the specific DF helped them. Afterward, we disclose the actual market value and ask for feedback again. In the first focus group, one participant criticized the difficult interpretation of state-of-the-art feature importance: "I was more bothered by the logic of the FI" (Alpha). Another participant mentioned the current complexity of the tool when enabling multiple XAI techniques: "When multiple explanations are shown, they can be independently confusing" (Gamma). In the second focus group, participants strongly highlighted the need for more global explanations of the AI, i.e., information about the underlying data as they stated: "Show summary statistics to the user and examples of correct (and incorrect) predictions" (Alpha). Additionally, one participant mentioned that we should "enable comparison options for the interactive feature, so the user does not have to remember the values of the last attempts" (Delta). Both focus groups highlighted the potential of the prototype and that it enables them to assess the quality of the AI advice. Overall, we could demonstrate the validity of our idea, the DPs, DFs, and instantiation. Furthermore, we could collect data for further refinement.

In this prototype paper, we propose design knowledge for resilient robo-advisors and build an artifact for house price appraisal for prospective real estate buyers and sellers. Based on existing literature, we derive design principles and instantiate them in the form of a configurable artifact. The feedback collected from two exploratory focus group workshops substantiated the utility of the artifact. With our work, we contribute to research and practice by deriving design knowledge as well as developing a unique artifact for real estate appraisal. In future work, we will incorporate the qualitative feedback, conduct additional confirmatory focus groups [15], and explore the utility of the artifact quantitatively in the form of a large-scale user experiment.

References

1. What is an ibuyer? when, why and how to sell your home to an ibuyer: Zillow, https://www.zillow.com/sellers-guide/what-is-an-ibuyer/, accessed: January 15, 2022
2. van der Aalst, W.M., Hinz, O., Weinhardt, C.: Resilient digital twins. Business & Information Systems Engineering **63**(6), 615–619 (2021)
3. Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M.T., Weld, D.: Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. pp. 1–16 (2021)
4. Buçinca, Z., Malaya, M.B., Gajos, K.Z.: To trust or to think: cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. Proceedings of the ACM on Human-Computer Interaction **5**(CSCW1), 1–21 (2021)
5. Burkart, N., Huber, M.F.: A survey on the explainability of supervised machine learning. Journal of Artificial Intelligence Research **70**, 245–317 (2021)
6. Cai, C.J., Jongejan, J., Holbrook, J.: The effects of example-based explanations in a machine learning interface. In: Proceedings of the 24th International Conference on Intelligent User Interfaces. pp. 258–262 (2019)
7. Gandel, S.: Zillow, facing big losses, quits flipping houses and will lay off a quarter of its staff., https://www.nytimes.com/2021/11/02/business/zillow-q3-earnings-home-flipping-ibuying.html, accessed: January 18, 2022
8. Gregor, S., Chandra Kruse, L., Seidel, S.: Research perspectives: the anatomy of a design principle. Journal of the Association for Information Systems **21**(6), 2 (2020)
9. Kucklick, J.P., Müller, J., Beverungen, D., Müller, O.: Quantifying the impact of location data for real estate appraisal - a gis-based deep learning approach. In: Proceedings of the 29th European Conference on Information Systems (2021)
10. Lyell, D., Magrabi, F., Raban, M.Z., Pont, L.G., Baysari, M.T., Day, R.O., Coiera, E.: Automation bias in electronic prescribing. BMC Medical Informatics and Decision Making **17**(1), 1–10 (2017)
11. Philippakis, A.S.: Structured what if analysis in dss models. In: Proceedings of the Hawaii International Conference on System Sciences. vol. 3, pp. 366–370 (1988)
12. Ribeiro, M.T., Singh, S., Guestrin, C.: "why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1135–1144 (2016)
13. Sidney Kess, J.: Selling and (perhaps) buying a home. The CPA Journal **86**(9), 54 (2016)
14. Suri, H.: Purposeful sampling in qualitative research synthesis. Qualitative Research Journal **11**(2), 63–75 (2011)
15. Tremblay, M.C., Hevner, A.R., Berndt, D.J.: Focus groups for artifact refinement and evaluation in design research. Communications of the Association for Information Systems **26**(1), 27 (2010)
16. Wexler, J., Pushkarna, M., Bolukbasi, T., Wattenberg, M., Viégas, F., Wilson, J.: The what-if tool: Interactive probing of machine learning models. IEEE transactions on visualization and computer graphics **26**(1), 56–65 (2019)
17. Zhang, Y., Liao, Q.V., Bellamy, R.K.: Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. pp. 295–305 (2020)