Manipulating and Measuring Model Interpretability

Forough Poursabzi-Sangdeh
forough.poursabzi@microsoft.com
Microsoft Research

Daniel G. Goldstein
dgg@microsoft.com
Microsoft Research

Jake M. Hofman
jmh@microsoft.com
Microsoft Research

Jennifer Wortman Vaughan
jenn@microsoft.com
Microsoft Research

Hanna Wallach
wallach@microsoft.com
Microsoft Research

Abstract

With the increased use of machine learning in decision-making scenarios, there has been a growing interest in creating human-interpretable machine learning models. While many such models have been proposed, there have been relatively few experimental studies of whether these models achieve their intended effects, such as encouraging people to follow the model's predictions when the model is correct and to deviate when it makes a mistake. We present a series of randomized, pre-registered experiments comprising 3,800 participants in which people were shown functionally identical models that varied only in two factors thought to influence interpretability: the number of input features and the model transparency (clear or black-box). Predictably, participants who were shown a clear model with a small number of features were better able to simulate the model's predictions. However, contrary to what one might expect when manipulating interpretability, we found no improvements in the degree to which participants followed the model's predictions when it was beneficial to do so. Even more surprisingly, increased transparency hampered people's ability to detect when the model makes a sizable mistake and correct for it, seemingly due to information overload. These counterintuitive results suggest that decision scientists creating interpretable models should harbor a healthy skepticism of their intuitions and empirically verify that interpretable models achieve their intended effects.

1 Introduction

Machine learning models are used to make decisions in critical domains like medical diagnosis [39], credit risk assessment [33], judicial sentencing and bail [3, 14, 35, 37], and hiring [43]. In addition, these models influence decisions about what news people see on social media platforms (e.g., Twitter and Facebook, [2, 10, 48, 61]), provide recommendations for what movies to watch (e.g., Netflix, [7]) or what music to listen to on online streaming platforms (e.g., Spotify, [54]), help people shop for clothing (e.g., Stitch Fix, [15]), and recommend whom to date on online dating platforms (e.g., OKCupid, [64]). More and more often, humans and machines decide jointly where humans used to decide alone. Because of the impact that model-aided decision-making has on business and policy outcomes, the importance of these interactions has been emphasized in the management science literature for decades [9, 47].

Machine learning models are often evaluated based on their predictive performance on held-out data sets, measured, for example, in terms of accuracy, precision, or recall. On these metrics, aggregated over many studies, machines often have

the advantage over people [31]. However, good performance on held-out data alone may not be sufficient to convince decision makers that a model is trustworthy, reliable, justifiable, ethical, or fair in practice. Historically, studies have shown that people are resistant to allowing decisions to be made entirely by algorithms [6, 18, 20, 60], though they can be supportive of using algorithmic decision aids provided they are first given the opportunity to review and potentially modify the predictions [21]. Recently, this resistance has been amplified by the difficulty of understanding how a potentially complex model has made a decision or how this decision can be improved. For example, it has been argued that certain black-box models are systematically unfair [27, 57, 70], sometimes causing managers to refrain from deploying them [17].

To address this problem as well as the General Data Protection Regulation's requirements to provide "meaningful information about the logic involved" in decisions made by algorithms, a new line of research has emerged that focuses on developing *interpretable* decision-making models. There are two common approaches. Since there is evidence that simple, transparent models can be as accurate as more complex, black-box models [4, 18, 28, 35, 65, 66], the first approach is to create inherently *simple* models, such as point systems that can be memorized [35, 69] or generative additive models in which the impact of each feature on the model's prediction should be relatively easy to understand [12, 50, 51]. The second is to provide post-hoc *explanations* for (potentially complex) models. One thread of research in this direction looks at how to explain individual predictions by learning simple local approximations of a model around particular data points [42, 52, 63] or estimating the influence of training examples [38], while another focuses on visualization of model output or properties [40, 71]. Both simple models and post-hoc explanations are attempts at creating more interpretable decisions aids.

But what is interpretability? Despite the progress in this area, there is still no consensus about how to define, quantify, or measure the interpretability of a machine learning model [24]. Different notions of interpretability, such as simulatability, trustworthiness, and simplicity, are often conflated [46]. This problem is exacerbated by the fact that there are different types of users of machine learning systems and these users may have different needs in different scenarios [67]. The approach that works best for a regulator who wants to understand why a particular person was denied a loan may be different from the approach that works best for a data scientist debugging a machine learning model or a manager using a model to make a business decision.

We take the perspective that the difficulty in assessing the value of interpretability stems from the fact that interpretability is not something that can be directly manipulated or measured. Rather, interpretability is a latent property that can be influenced by different *manipulable factors* (such as the number of features, the complexity of the model, the transparency of the model, or even the user interface) and that impacts different *measurable outcomes* (such as people's tendency to follow the model's predictions or people's ability to simulate or debug the model's predictions). Different factors may influence these outcomes in different ways. As such, we argue that to understand interpretability, it is necessary to directly manipulate and measure the influence that different factors have on people's abilities to complete various tasks. What is or is not "interpretable" should be defined by user behavior, not by what appeals to the modeler's intuition [55, 56].

We build on decision-making research on human trust in algorithmic (also called actuarial or mechanical) decision making [19–21, 49, 58, 60, 73] and similar efforts in computer science [8, 26, 30, 41, 44, 45, 62, 63] and present a sequence of pre-registered experiments ($N = 3,800$) in which we vary factors thought to make models more or less interpretable and measure how these changes impact people's decision making. Based on a structured review of the literature on effective decision aids and interpretable machine learning, we focus on two factors that are often assumed to influence interpretability, but rarely studied formally—the number of features (i.e., predictors) and the model transparency (i.e., whether the model internals are *clear* or a *black box*)—and investigate how these factors affect three outcomes commonly mentioned in the literature: simulatability, deviation, and error detection. Specifically, we look at the following questions:

- **How well can people estimate what a model will predict?** Here we find that people can better simulate a model's predictions when presented with a clear model with few features compared to the other experimental conditions.
- **To what extent do people follow a model's predictions when it is beneficial to do so?** Contrary to what one might expect when manipulating factors associated with interpretability, we do not find a significant improvement in the degree to which people follow the predictions of a clear model with few features compared to the other experimental conditions.
- **How well can people detect when a model has made a sizable mistake?** Even more surprisingly, we find that transparency can hamper people's ability to detect when a model makes serious mistakes.

In each of our experiments, participants were asked to predict the prices of apartments in a single neighborhood in New York City with the help of a machine learning model. We conducted our experiments on laypeople, as they represent a large class of individuals who might potentially use or be impacted by machine learning models, and chose the task of predicting housing prices because many people have considered purchasing a house or apartment, making the setting both familiar and potentially interesting to participants. Each apartment was represented in terms of eight features: number of bedrooms, number of bathrooms, square footage, total rooms, days on the market, maintenance fee, distance from the subway, and distance from a school. All participants saw the same set of apartments (i.e., the same feature values) and, crucially, the same model prediction for each apartment, which came from a linear regression model. What varied between the experimental conditions was *only the presentation of the model*. As a result, any observed differences in the participants' behavior between conditions could be attributed entirely to the model presentation. This is a key feature of our experimental design.

In our first experiment (Section 2), we tested people on a sequence of 12 apartments. The first 10 apartments had typical configurations whereas the last two had unusual combinations of features (more bathrooms than bedrooms). For each apartment, participants first saw its features alongside the model and were asked to estimate what the model would predict for the apartment's selling price. They were then shown the model's prediction and asked for their own estimate of what the apartment sold for.

We hypothesized that participants who were shown a clear model with a small number of features would be better able to simulate the model's predictions and be less likely to deviate from (that is, more likely to follow) the model's predictions. We also hypothesized that the assigned conditions would affect participants' abilities to correct the model's inaccurate predictions about the unusual apartments. As predicted, participants who were shown a clear model with a small number of features were better able to simulate the model's predictions. However, we did not find that people followed the clear, simple model more closely than the complex, opaque model. This is of concern because people's predictions were less accurate than the model's, a familiar finding in the literature on human versus statistical prediction [19, 31]. Even more surprisingly, participants who were shown a clear model were less likely to correct the model's inaccurate predictions on the unusual apartments. To account for these unexpected results, we ran three additional experiments.

In our second experiment (Section 3), we scaled down the apartment prices and maintenance fees to match median housing prices in the U.S. in order to determine whether the findings from our first experiment were merely an artifact of New York City's high prices. With scaled-down prices and fees, the findings from our first experiment replicated quite closely.

Our third experiment (Section 4) was designed to measure *weight of advice*—commonly used in the literature on advice-taking [29, 72] and subsequently used in the context of algorithmic predictions by Logg [49]—to quantify the extent to which people update their belief towards advice they are given. We calculated weight of advice by eliciting two predictions from participants: one before being introduced to the model and one after. Similar to deviation, we found

no meaningful difference in weight of advice between the clear, two-feature model and the black-box, eight-feature model. Interestingly, however, participants in the clear conditions were better able to correct the model's mistakes than in the previous two experiments. One possible reason for this improvement is that making predictions before being exposed to the clear model's overwhelming level of detail increased the likelihood that participants noticed unusual combinations of features and adjusted their estimates accordingly.

This motivated our fourth and final experiment (Section 5), which focused on whether the observed differences in error detection between conditions could be explained by people in the clear conditions failing to notice the configurations of the unusual apartments due to information overload. We returned to the setup from the first experiment but added an "outlier focus" condition in which some people were shown additional text highlighting apartments with peculiar configurations as outliers. Without this manipulation we again found that participants who were shown a clear model were less likely to correct inaccurate predictions on examples for which the model made a sizable mistake. However, this difference essentially disappeared for participants in the outlier focus condition, consistent with the idea that transparency can be overwhelming and cause users to overlook unusual cases. Seeing this effect, we revisited the results of our first two experiments with an eye towards information overload in a post-hoc (i.e., not pre-registered) analysis. We found that the condition that provides users with the most information (CLEAR-8) scored worst on simulation error, deviation, and prediction error compared to the other conditions, suggesting that it may be detrimental to present people with models that are both transparent and complex. We present these results in the discussions of Experiments 1 and 2.

In summary, across several experiments involving several thousand participants we saw that factors thought to improve interpretability often had negligible effects and even had detrimental effects in certain instances. While one might expect that exposing model internals by default could only help decision makers, our experiments suggest otherwise. Taken together, these results emphasize the importance of user testing over intuition in the design of interpretable models to assist decision making.

In the remainder of the paper we provide further details for each of these experiments and present our results in greater depth. We conclude by discussing limitations of and possible extensions to our work, as well as implications for the design of interfaces to decision aids.

2   Experiment 1: Predicting prices

Our first experiment was designed to measure the influence of the number of features and model transparency (clear or black box) on three behaviors that our literature search revealed to be commonly associated with interpretability: laypeople's abilities to predict (or simulate) the model's predictions, follow a model's prediction when it would be beneficial to do so, and detect when a model makes mistakes. Before running the experiment, we posited and pre-registered three hypotheses, stated informally here:[1]

H1. **Simulation.** A clear model with a small number of features will be easiest for participants to simulate.
H2. **Deviation.** On typical examples participants will be more likely to follow (that is, less likely to deviate from) the predictions of a clear model with a small number of features than the predictions of a black-box model with a large number of features.
H3. **Detection of mistakes.** Participants in different conditions will exhibit varying abilities to correct the model's inaccurate predictions on unusual examples.

---

[1]Pre-registered hypotheses are available at `https://aspredicted.org/xy5s6.pdf`.

We test the first hypothesis by showing people an apartment configuration, asking them to estimate what the model will predict for its selling price, and comparing this estimate with the model's prediction. A small difference between these two quantities would indicate that people have a good understanding of how the model works. For the second hypothesis, we quantify the degree to which participants follow a model by showing them the model's prediction for each apartment, asking them to estimate the actual selling price, and measuring the deviation between this estimate and the model's prediction.

We use the same deviation measure for the third hypothesis, but applied to unusual apartments for which the model makes erroneously high predictions. Here, a large deviation from the model's predictions would imply that people are able to correct the model's mistakes.

For the unusual examples, we did not make directional predictions about which conditions would make participants more or less able to correct inaccurate predictions. On the one hand, if a participant understands the model better, she may be better equipped to correct examples on which the model makes mistakes. On the other hand, a participant may follow a model more if she understands how it arrives at its predictions.

We additionally pre-registered our intent to analyze participants' prediction error in each condition, but intentionally refrained from making directional predictions.

2.1 Experimental design

As explained in the previous section, we asked participants to predict apartment prices with the help of a machine learning model. We manipulated the presentation of the model in a 2×2 design:

- Participants were randomly assigned to see either a model that uses only two features (number of bathrooms and square footage—the two most predictive features) or a model that uses all eight features.
- Participants were randomly assigned to either see the model internals (i.e., a linear regression model with visible coefficients) or see the model as a black box.

We showed all participants the same set of apartments and the same model prediction for each apartment, regardless of their randomly assigned experimental condition. This key feature of our experimental design is what enabled us to run tightly controlled experiments. *The only thing that varied between the primary conditions was model presentation.*

Screenshots from each of the four primary experimental conditions are shown in Figure 1. Note that all eight feature values were visible to participants in all conditions. Furthermore, the accuracies of the two- and eight-feature models were nearly identical, as described below, and their (rounded) predictions were identical on all apartments used in the experiment (see Appendix .2). We additionally included a baseline condition in which there was no model available.

We ran the experiment on Amazon Mechanical Turk using psiTurk [32], an open-source platform for designing online experiments and the experiment was IRB-approved. We recruited 1,250 participants, all located in the U.S., with Mechanical Turk approval ratings greater than 97%.[2] The participants were randomly assigned to the five conditions (CLEAR-2, $n = 248$; CLEAR-8, $n = 247$; BB-2, $n = 247$; BB-8, $n = 256$; and NO-MODEL, $n = 252$) and each participant received a flat payment of \$2.50.

---

[2]Multiple studies show that data from high reputation Mechanical Turk workers are comparable to data from commercial panels and university pools when assessing factors such as attentiveness, honesty, and effort [11, 13, 16, 53, 59].

(a) Clear, two-feature condition (CLEAR-2).

(b) Black-box, two-feature condition (BB-2).

(c) Clear, eight-feature condition (CLEAR-8).

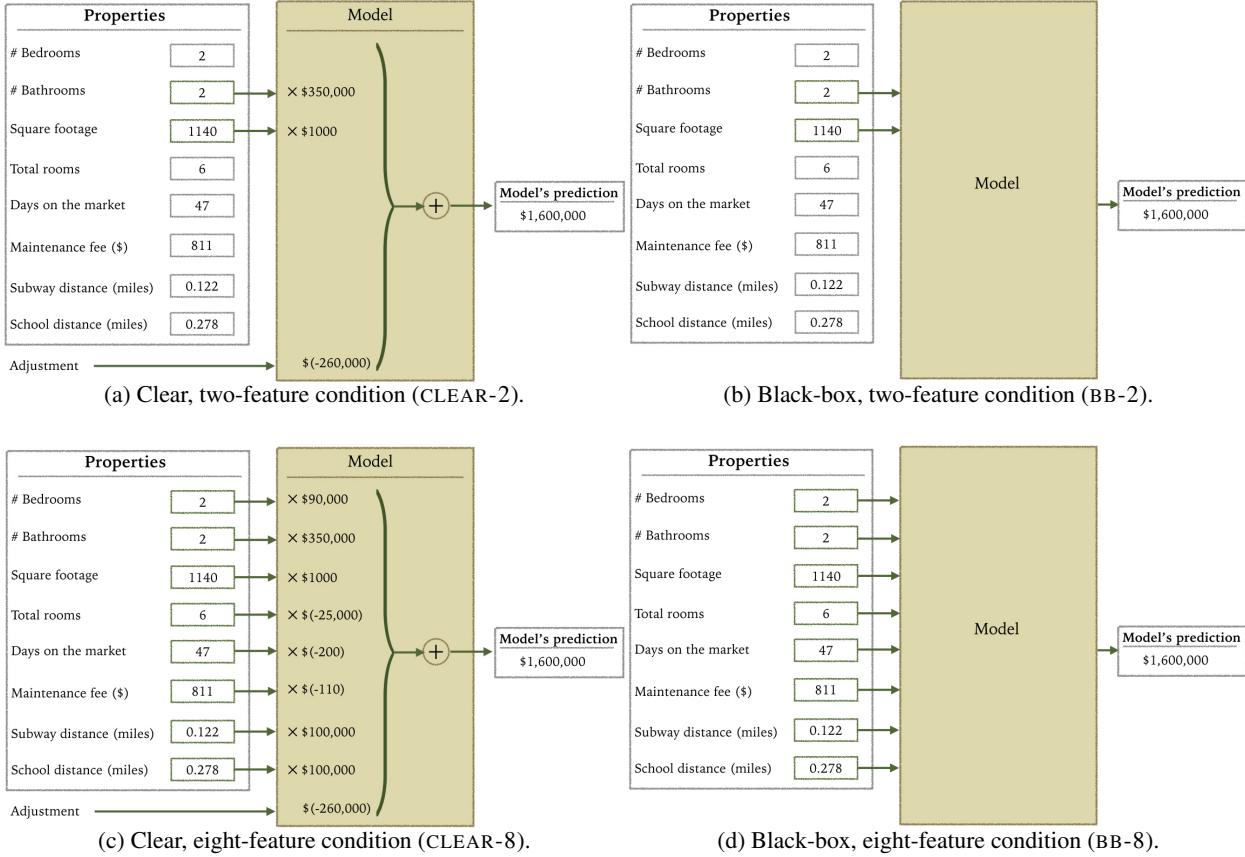(d) Black-box, eight-feature condition (BB-8).

Figure 1: The four primary experimental conditions. In the conditions on top, the model used two features; on the bottom, it used eight. In the conditions on the left, participants saw the model internals; on the right, they were presented with the model as a black box.

Participants were first shown detailed instructions, including, in the clear conditions, a simple English description of the corresponding two- or eight-feature linear regression model (Appendix .1). To be sure they understood these instructions, participants were required to answer a multiple choice question on the number of features used by the model before proceeding with the experiment in two phases. The *training phase* familiarized participants with both the housing domain and the model's predictions. Participants were shown ten apartments in a random order. In the four primary experimental conditions, participants were shown the model's prediction of each apartment's price, asked to make their own prediction, and then shown the apartment's actual price. In the baseline condition, participants were asked to predict the price of each apartment and then shown the actual price.

In the *testing phase*, participants were shown twelve apartments they had not previously seen. The order of the first ten was randomized, while the remaining two always appeared last, for reasons described below. In the four primary experimental conditions, participants were asked to guess what the model would predict for each apartment (i.e., simulate the model) and to indicate how confident they were in this guess on a five-point scale (Figure 2a). They were

What do you think the model will predict?

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2 2.1 2.2 2.3 2.4 2.5 2.6 2.7 2.8 2.9 3

$800,000

How confident are you the model will predict this?

	1	2	3	4	5	
It's likely the model will predict something else	○	○	⦿	○	○	I'm confident the model will predict this

(a) Step 1: Participants were asked to guess the model's prediction and state their confidence.

What you thought the model would predict:

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2 2.1 2.2 2.3 2.4 2.5 2.6 2.7 2.8 2.9 3

$800,000

What the model actually predicted:

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2 2.1 2.2 2.3 2.4 2.5 2.6 2.7 2.8 2.9 3

$600,000

How confident are you that the model got it right?

	1	2	3	4	5	
I'm confident the model got it wrong	○	○	○	⦿	○	I'm confident the model got it right

(b) Step 2: Participants were asked to state their confidence in the model's prediction.

What you thought the model would predict:

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2 2.1 2.2 2.3 2.4 2.5 2.6 2.7 2.8 2.9 3

$800,000

What the model actually predicted:

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2 2.1 2.2 2.3 2.4 2.5 2.6 2.7 2.8 2.9 3

$600,000

What do you think this apartment actually sold for?

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2 2.1 2.2 2.3 2.4 2.5 2.6 2.7 2.8 2.9 3

$700,000

How confident are you that you got it right?

	1	2	3	4	5	
It's likely I got it wrong	○	○	○	⦿	○	I'm confident I got it right

(c) Step 3: Participants were asked to make their own prediction and state their confidence.

Figure 2: Part of the testing phase in the first experiment.

then shown the model's prediction and asked to indicate how confident they were that the model was correct (Figure 2b). Finally, they were asked to make their own prediction of the apartment's price and to indicate how confident they were in this prediction (Figure 2c). In the baseline condition, participants were asked to predict the price of each apartment and to indicate their confidence.

The apartments shown to participants were selected from a data set of apartments sold between 2013 and 2015 on the Upper West Side of New York City, and were taken from StreetEasy.com, a popular real estate website. To create the models for the four primary experimental conditions, we first trained a two-feature linear regression model on our data set, rounding coefficients for readability.[3] To keep the models as similar as possible, we fixed the coefficients for number of bathrooms and square footage and the intercept of the eight-feature model to match those of the two-feature model, and then trained a linear regression model with the remaining six features, following the same rounding procedure. The models explain 82% of the variance in the apartment prices. When presenting the model predictions to participants, we rounded predictions to the nearest \$100,000. The model coefficients are shown in Figure 1.

We used the same set of apartments for each participant since randomizing the selection of apartments would introduce additional noise and reduce the power of the experiments, making it harder to spot differences between conditions. To enable comparisons across experimental conditions, the ten apartments used in the training phase and the first ten apartments used in the testing phase were selected from those in our data set for which the rounded predictions of the two- and eight-feature models agreed. They were chosen to cover a representative range of the models' prediction errors. Details on the apartment configurations and how they were selected are provided in Appendix .2.

The last two apartments used in the testing phase were chosen to test our third hypothesis, namely, that participants will differ in how much they deviate from the model's inaccurate predictions on unusual examples. To test this hypothesis, we would ideally have used an apartment with strange or misleading features that caused the two- and eight-feature models to make the same bad prediction. Unfortunately, there was no such apartment in our data set, so we chose two examples to test different aspects of our hypothesis. Both of these examples exploited the models' large coefficient (\$350,000) for the number of bathrooms. The first (apartment 11) was a one-bedroom, two-bathroom apartment from our data set for which both models made high, but different, predictions. Comparisons between the two- and eight-feature conditions were therefore impossible, but we could examine differences in prediction error between the clear and black-box conditions. The second (apartment 12) was a synthetically generated one-bedroom, three-bathroom apartment for which both models made the same (high) prediction, allowing comparisons between all conditions, but ruling out prediction error comparisons since there was no ground truth. These apartments were always shown last to avoid the phenomenon in which people trust a model less after seeing it make a mistake [20].

2.2   Results

---

[3]For each estimated coefficient, we found a round number that was within one quarter of a standard error of the estimate.

Experiment 1: New York City prices

(a)   (b)
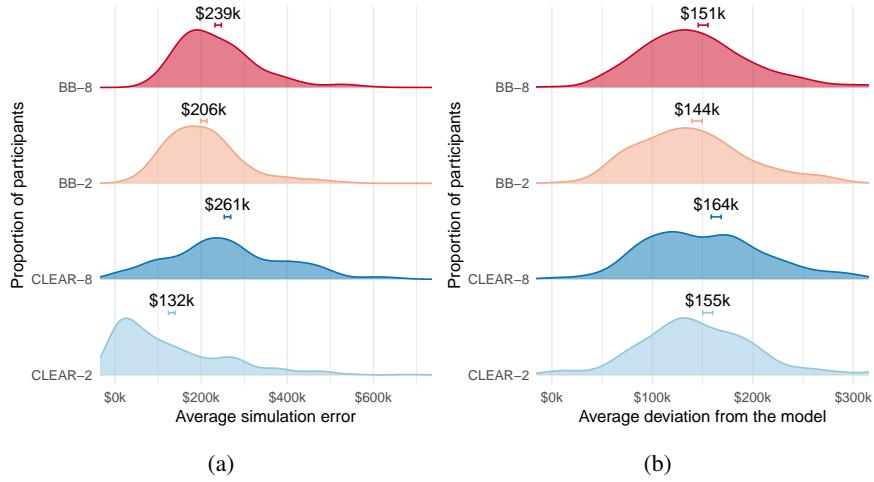
Figure 3: Results from Experiment 1: density plots for participants' (a) mean simulation error and (b) mean deviation from the model's predictions. Numbers in each subplot indicate average values over all participants in the corresponding condition and error bars indicate one standard error.

Experiment 2: Representative U.S. prices

(a)   (b)
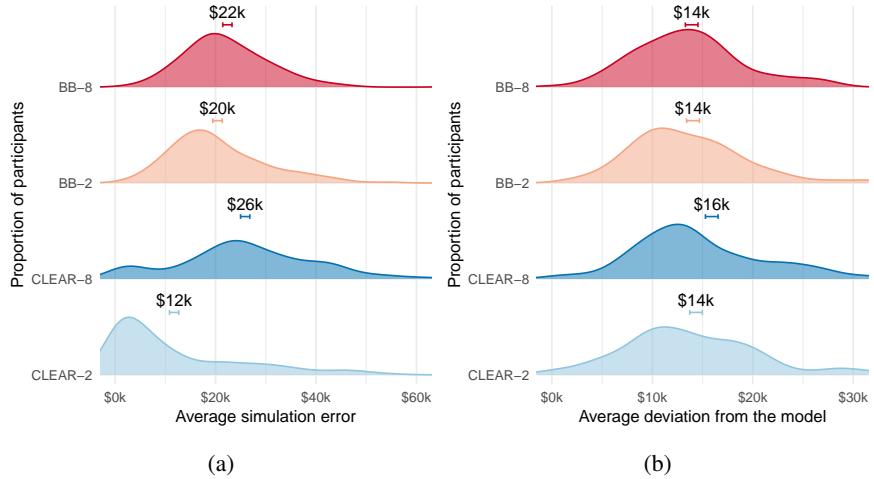
Figure 4: Results from Experiment 2, which replicate the findings of Experiment 1.

Having run our experiment, we compared participants' behavior across the conditions.[4] Doing so required us to compare multiple responses from multiple participants, which was complicated by possible correlations among any given participant's responses. For example, some people might consistently overestimate apartment prices regardless of the condition they are assigned to, while others might consistently provide underestimates. We addressed this by fitting a mixed-effects model for each measure of interest to capture differences across conditions while controlling for participant-level effects—a standard approach for analyzing repeated measures experimental designs [5]. We derived all statistical tests from these models. Bar plots and the mean outcomes in the density plots show averages ($\pm$ one standard error) by condition from the fitted models. To test for our specific hypotheses, we ran contrasts and report degrees of freedom, test statistics, and p-values under these models. Unless otherwise noted, all plots and statistical tests correspond to just the first ten apartments from the testing phase. The results corresponding to the pre-registered hypotheses are as follows:

H1. **Simulation.** We defined a participant's simulation error to be $|m - u_m|$, the absolute deviation between the model's prediction m and the participant's guess for that prediction u_m. Figure 3a shows the simulation error in the testing phase. Participants in the CLEAR-2 condition had lower simulation error, on average, than participants in the other conditions ($t(994) = -12.06$, $p < 0.001$). This indicates that, as hypothesized, participants who saw a clear model with a small number of features were better able to predict what the model would predict.

H2. **Deviation.** We quantified a participant's deviation from the model as $|m - u_a|$, the absolute difference between the model's prediction m and the participant's prediction of the apartment's price u_a. Smaller values indicate that participants follow the model's predictions more closely. Figure 3b shows that contrary to our second hypothesis, we found no significant difference in participants' deviation from the model between CLEAR-2 and BB-8 ($t(994) = 0.67$, $p = 0.5$).

H3. **Detection of mistakes.** We used the last two apartments in the testing phase (apartments 11 and 12) to test our third hypothesis. The models err and make overly high predictions on these examples because they do not consider the ratio of the number of bathrooms to the number of bedrooms. For both apartments participants in the four primary conditions overestimated the apartments' prices compared to participants in the baseline condition. We suspect that this is due to an anchoring effect around the models' predictions. For apartment 11, we found no significant difference in participants' deviation from the model's prediction between the four primary conditions ($F(3, 994) = 1.03$, $p = 0.379$ under a one-way ANOVA). For apartment 12, a one-way ANOVA revealed a significant difference between the four primary conditions ($F(3, 994) = 4.42$, $p = 0.004$). Participants in the clear conditions deviated from the model's prediction less, on average, than participants in the black-box conditions ($F(1, 994) = 8.81$, $p = 0.003$ for the main effect of model transparency, see Figure 5a), resulting in even worse final predictions of the apartment's price. Participants' poor abilities to detect the model's mistakes in clear conditions contradicts the common intuition that transparency enables users to detect when a model is making a mistake. We explore this finding in more detail later.

We additionally performed some post-hoc analyses. First, we checked participants' self-reported confidence in the model's predictions. Though this was not a primary outcome that we pre-registered a hypothesis for, it showed an interesting difference between participants' stated confidence and revealed behavior. Specifically, even though participants reported that they were more confident in the predictions of the model in the CLEAR-2 condition than in the BB-8 condition (a difference of .25 on average on a five-point scale from "I'm confident the model got it wrong" to "I'm confident the model got it right", $t(994) = 4.27$, $p < 0.001$), their decisions did not reflect this. Predictions in

---

[4]For all of our experiments, we report all sample sizes, conditions, data exclusions, and measures for the main analyses mentioned in the pre-registration documents. We determined the sample size for our first experiment based on estimates from a small pilot experiment to enable detecting a difference of at least $50,000 in deviation between the CLEAR-2 and BB-8 conditions with 80% power. For the subsequent experiments, we adjusted the sample size to target a power of 80% or more. Full distributions of collected responses are in Appendix .3 and details of statistical tests are in Appendix .4. All data and code will be made available upon publication.
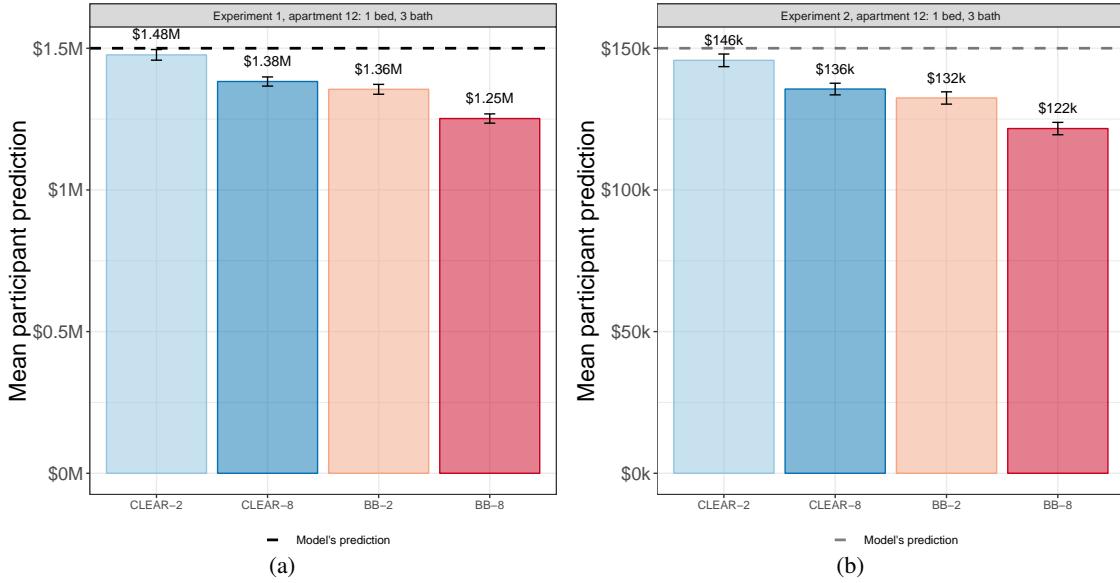
Figure 5: Mean predictions of the prices of apartment 12 in (a) Experiment 1 and (b) Experiment 2. Error bars indicate one standard error.

both conditions deviated from the model by the same amount, on average.

The second post-hoc analysis concerned prediction error, which we measured as $|a - u_a|$, the absolute difference between the apartment's actual price a, and the participant's prediction of the apartment's price, u_a. Figure 6a shows that on apartments with a normal configuration, participants in the primary conditions had higher average error than the model itself, but fared better than those in the baseline condition ($t(1245) = 15.28$, $p < 0.001$ for the comparison of the baseline with the four primary conditions), indicating that the use of a model is highly advantageous in this setting. On the other hand, participants' prediction error on apartment 11 revealed a reversed pattern: participants in the primary conditions had lower prediction error than the model but higher prediction error than those in the baseline condition ($t(1245) = -7.99$, $p < 0.001$). Having a model helped for typical apartments but was detrimental on an apartment with an unusual configuration for which the model made a highly inaccurate prediction. Transparency further hampered participants' prediction quality: participants who were shown a clear model made worse predictions of the apartment's selling price than those who were shown a black-box model ($F(1, 994) = 31.98$, $p < 0.001$ for the main effect of model transparency under a two-way ANOVA).

As predicted, we found that participants who were shown a clear model with a small number of features were better able to simulate the model's predictions. However, we did not find that they followed the model's predictions more closely when it would have been beneficial to do so. We also found that, contrary to intuition, participants who were shown a clear model were less able to correct the model when it made a mistake. Finally, we found no meaningful difference in prediction error between the four primary conditions. We found that participants who had the help of a model were more accurate than those in the baseline condition who did not. Having a model mattered but participants would have been better off simply following the model, on average, than adjusting its predictions.
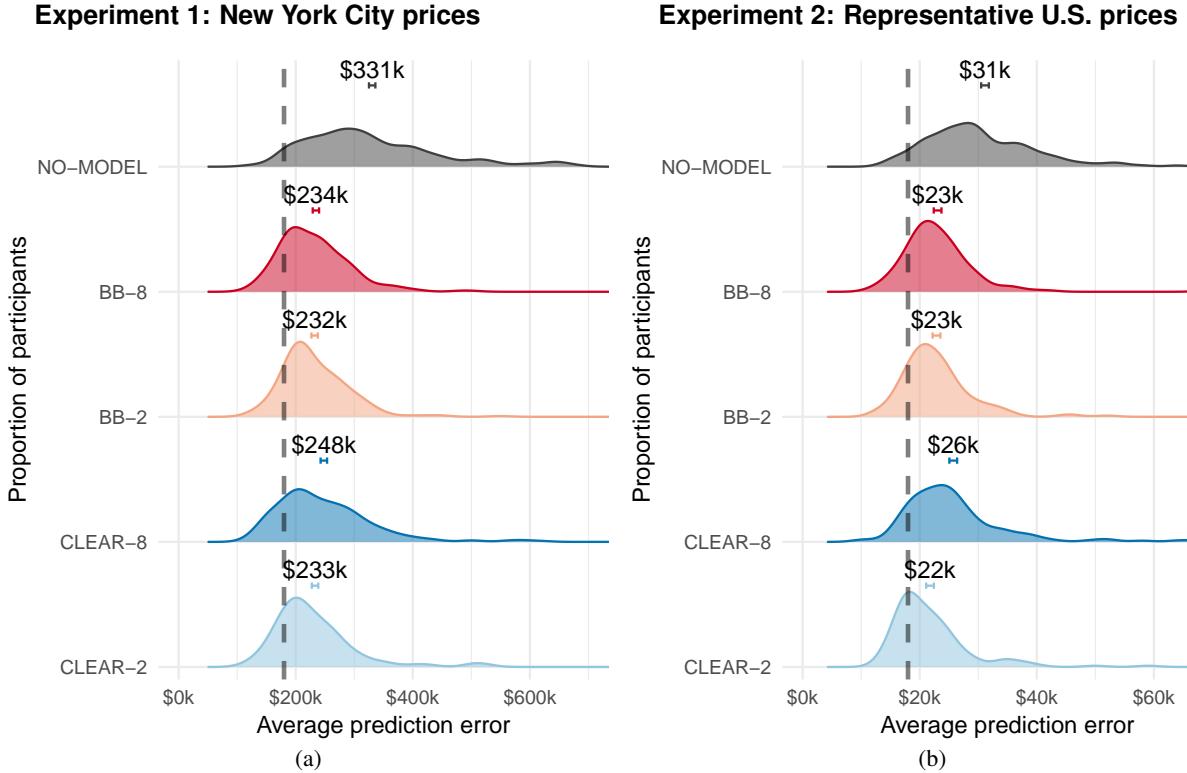
Figure 6: Density plots for participants' mean prediction error in Experiment 1 (left) and in Experiment 2 (right). Numbers in each subplot indicate average values over all participants in the corresponding condition and error bars indicate one standard error. Vertical lines indicate the mean prediction error of the model on the first ten apartments in Experiment 1 and Experiment 2.

3 Experiment 2: Scaled-down prices

One potential critique of our first experiment is that participants' lack of familiarity with New York City's unusually high apartment prices may influence their willingness to follow the model and ability to detect when the model makes a mistake. Our second experiment was designed as a robustness check to address this potential concern by replicating our first experiment with apartment prices and maintenance fees scaled down to match median housing prices in the U.S. Before running this experiment we pre-registered three hypotheses. The first two hypotheses (H4 and H5) are identical to H1 and H2 from the first experiment. We made the third hypothesis (H6) more precise than H3 to reflect the results of the first experiment and the results of a small pilot with scaled-down prices:

H6. **Detection of mistakes.** Participants will be less likely to correct inaccurate predictions on unusual examples of a clear model compared to a black-box model, and this effect will be more prominent the more unusual an example is.

3.1 Experimental design

We first scaled down the apartment prices and maintenance fees from our first experiment by a factor of ten. To account for this change, we also scaled down all regression coefficients (except for the coefficient for maintenance fee) by a factor of ten. Apart from the description of the neighborhood from which the apartments were selected, the experimental design was unchanged. We again ran the experiment on Amazon Mechanical Turk. We excluded people who had participated in our first experiment, and recruited 750 new participants all of whom satisfied the selection criteria from the first experiment. The participants were randomly assigned to the five conditions (CLEAR-2, $n = 150$; CLEAR-8, $n = 150$; BB-2, $n = 147$; BB-8, $n = 151$; and NO-MODEL, $n = 152$) and each participant received a flat payment of $2.50.

3.2 Results

The results of this experiment closely replicated those of Experiment 1.

H4. **Simulation.** As hypothesized, and shown in Figure 4a, participants in the CLEAR-2 condition had significantly lower simulation error, on average, than participants in the other primary conditions ($t(594) = -10.41, p < 0.001$). This is in line with the finding from the first experiment.

H5. **Deviation.** In line with the first experiment, and contrary to our original hypothesis, we found no significant difference in participants' deviation from the model on typical apartments between CLEAR-2 and BB-8 ($t(594) = 0.49$, $p = 0.626$, see Figure 4b).

H6. **Detection of mistakes.** For apartment 12, as hypothesized and in line with our findings from Experiment 1, a one-way ANOVA revealed a significant difference across the four primary conditions ($F(3, 594) = 7.96, p < 0.001$). In particular, participants in the clear conditions deviated from the model's prediction less, on average, than participants in the black-box conditions for apartment 12, resulting in even worse final predictions of the apartment's price ($t(594) = -4.16, p < 0.001$, see Figure 5b).

For apartment 11, while a one-way ANOVA revealed a significant difference across the four primary conditions ($F(3, 594) = 3.00$, $p = 0.03$), as in Experiment 1, we found no significant difference in participants' deviation from the model's prediction between the clear and black-box conditions ($t(594) = -1.82, p = 0.069$), perhaps because the configuration is not sufficiently unusual. These findings both suggest that New York City's high apartment prices do not explain participants' poor abilities to correct inaccurate predictions.

As in Experiment 1, we conducted some post-hoc analyses. For self-reported confidence, unlike in Experiment 1, we found no significant difference between CLEAR-2 and BB-8 ($t(594) = 1.03, p = 0.303$). We note that the effect size of this difference in Experiment 1 was small (Cohen's d of 0.23) and even smaller in Experiment 2 (Cohen's d of 0.07), which was a nearly identical experiment except for the prices. We also note that despite the difference in self-reported confidence, there was no significant difference in deviation from the model in either experiment.

Results for prediction error are in line with those in Experiment 1. For the typical apartments, participants in the primary conditions had higher average error than the model itself, but fared better than participants in the baseline condition ($t(745) = 10.62, p < 0.001$). A one-way ANOVA revealed a significant difference in prediction error across the four primary conditions ($F(3, 594) = 8.60, p < 0.001$). Figure 6b shows that the prediction error in the CLEAR-8 condition is particularly high. Looking at the results from Experiment 1, we see a similar pattern: prediction error in the

CLEAR-8 condition was higher than in the other three primary conditions ($t(994) = 2.68$, $p = 0.007$ for Experiment 1, $t(594) = 4.78$, $p < 0.001$ for Experiment 2, see Figure 6).

Contrary to the case with typical apartments, having a model was detrimental for predicting the price of the atypical apartment 11. Participants in the primary conditions made better predictions than the model but worse predictions than those in the baseline condition ($t(745) = -6.41$, $p < 0.001$). As in Experiment 1, transparency further hampered people's predictions on apartment 11: participants who were shown a clear model made worse predictions of the apartment's selling price than those who were shown a black-box model ($t(594) = 2.63$, $p = 0.008$ for the contrast of CLEAR-2 and CLEAR-8 with BB-2 and BB-8). These findings, coupled with the difficulty that people had in simulating the CLEAR-8 model compared to the other conditions ($t(994) = 7.96$, $p < 0.001$ for Experiment 1, $t(594) = 7.23$, $p < 0.001$ for Experiment 2, see Figures 3a and 4a), and the higher deviation of participants' predictions from the model's predictions in the CLEAR-8 condition compared to other conditions ($t(994) = 2.37$, $p = 0.018$ for Experiment 1, $t(594) = 2.49$, $p = 0.012$ for Experiment 2, see Figures 3b and 4b), suggest potential downsides of decision aids that are both transparent *and* complex.

Experiment 2 replicated the results of Experiment 1, suggesting that they were not an artifact of participants' lack of familiarity with New York City's high prices. Both experiments showed that participants who used a simple, transparent model were better able to simulate the model's predictions. However, they were not more inclined to follow these predictions, or to detect when the model was making a sizable mistake.

4   Experiment 3: Weight of Advice

In our first two experiments, participants were no more likely to follow the predictions of a clear model with a small number of features than the predictions of a black-box model with a large number of features, as indicated by the deviation of their own predictions from the model's prediction. However, perhaps another measure would reveal differences between the conditions. In this section, we therefore present our third experiment, which was designed to allow us to compare a measure frequently used in the literature on advice-taking (*weight of advice*, [29, 49, 72]) across the conditions.

Weight of advice quantifies the degree to which people update their beliefs (e.g., predictions made *before* seeing the model's predictions) toward advice they are given (e.g., the model's predictions). In the context of our experiment, it is defined as $|u_2 - u_1| / |m - u_1|$, where m is the model's prediction, u_1 is the participant's initial prediction of the apartment's price before seeing m, and u_2 is the participant's final prediction of the apartment's price after seeing m. It is equal to 1 if the participant's final prediction matches the model's prediction and equal to 0.5 if the participant averages their initial prediction and the model's prediction.

To understand the benefits of comparing weight of advice across the conditions, consider the scenario in which a participant's final prediction u_2 is close to the model's prediction m. There are different reasons why this might happen. On the one hand, it could be the case that the initial prediction u_1 was far from m and the participant made a significant update to their initial prediction based on the model. On the other hand, it could be the case that the initial prediction u_1 was already close to m and the participant did not update her prediction at all. These two scenarios are indistinguishable in terms of the participant's deviation from the model's prediction. In contrast, weight of advice would be high in the first case and low in the second.

We additionally used this experiment to check whether participants' behavior would differ if they were told that the
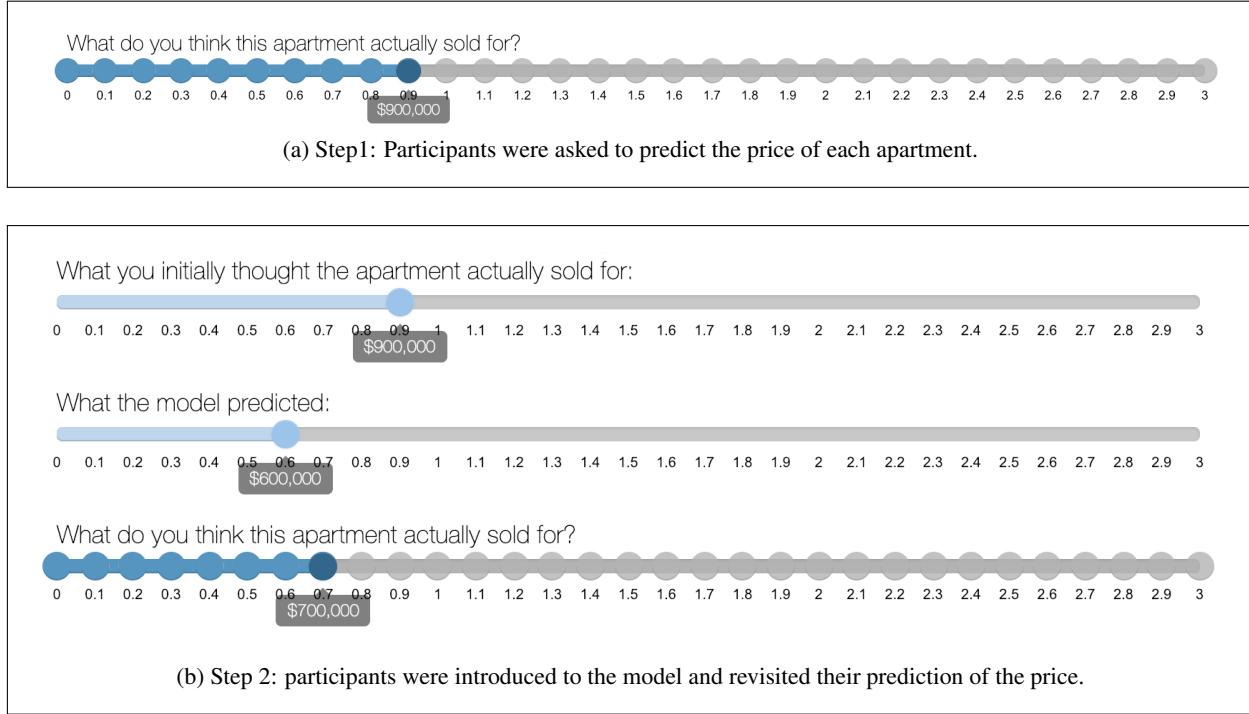
(a) Step1: Participants were asked to predict the price of each apartment.

(b) Step 2: participants were introduced to the model and revisited their prediction of the price.

Figure 7: Part of the testing phase in the third experiment.

predictions were made by a "human expert" instead of a model. Previous studies have examined this question from different perspectives with differing results [20, 22, 23, 25, 58]. Most closely related to our experiment is work by Logg [49], which found that when people have no information about the quality of predictions shown to them, they follow the predictions that appear to come from an algorithm more than those that appear to come from a human. We were curious to see if this finding extended to a setting where people were first given the chance to gauge the quality of predictions before deciding how closely to follow them. We pre-registered four hypotheses:[5]

H7. **Deviation.** Participants' predictions will deviate less from the predictions of a clear model with a small number of features than the predictions of a black-box model with a large number of features.

H8. **Weight of advice.** Weight of advice will be higher for participants who see a clear model with a small number of features than for those who see a black-box model with a large number of features.

H9. **Humans vs. machines.** Participants' deviation and weight of advice measures will differ depending on whether the predictions come from a black-box model with a large number of features or a human expert.

H10. **Detection of mistakes.** Participants in different conditions will exhibit varying abilities to correct the model's inaccurate predictions on unusual examples.

The first two hypotheses are variations on H2 from our first experiment, while the last hypothesis is identical to H3.

---

[5]Pre-registered hypotheses are available at `https://aspredicted.org/795du.pdf`.

4.1 Experimental design

We returned to using the original New York City housing prices and used the same four primary experimental conditions as in the first two experiments plus a new condition, EXPERT, in which participants saw the same information as in BB-8, but with the black-box model labeled as "Human Expert" instead of "Model." We did not include a baseline condition because the most natural baseline would have been to simply ask participants to predict apartment prices (i.e., the first step of the testing phase described below).

We again ran the experiment on Amazon Mechanical Turk. We excluded people who had participated in our first two experiments, and recruited 1,000 new participants all of whom satisfied the selection criteria from our first two experiments. The participants were randomly assigned to the five conditions (CLEAR-2, $n = 202$; CLEAR-8, $n = 200$; BB-2, $n = 202$; BB-8, $n = 198$; and EXPERT, $n = 197$) and each participant received a flat payment of \$1.50. We excluded data from one participant who reported technical difficulties.

We asked participants to predict apartment prices for the same set of apartments used in the first two experiments. However, in order to calculate weight of advice, we modified the experiment so that participants were asked for two predictions for each apartment during the testing phase: an initial prediction before being shown the model's prediction and a final prediction after being shown the model's prediction. To ensure that participants' initial predictions were not influenced by the condition they were assigned to, we asked for their initial predictions for all twelve apartments before introducing them to the model or human expert and before informing them that they would be able to update their predictions.[6]

Participants were first shown detailed instructions (which intentionally did not include any information about the corresponding model or human expert), before proceeding with the experiment in two phases. In the (short) training phase, participants were shown three apartments, asked to predict each apartment's price, and shown the apartment's actual price. The testing phase consisted of two steps. In the first step, participants were shown another twelve apartments. The order of all twelve apartments was randomized. Participants were asked to predict the price of each apartment (Figure 7a). In the second step, participants were introduced to the model or human expert before revisiting the twelve apartments (Figure 7b). As in the first two experiments, the order of the first ten apartments was randomized, while the remaining two (apartments 11 and 12) always appeared last. For each apartment, participants were first reminded of their initial prediction, next shown the model or expert's prediction, and only then asked to make their final prediction of the apartment's price. To simplify the experiment, we did not ask for participants' confidence in their and the model's predictions.
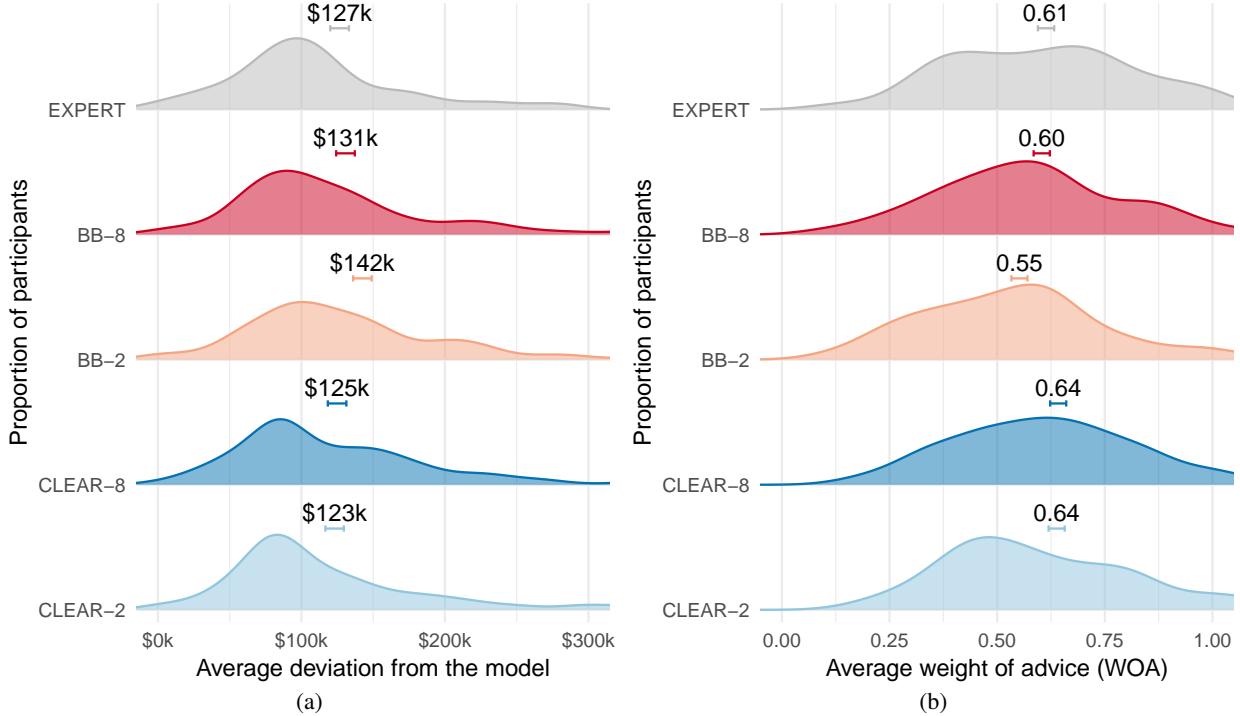
Figure 8: Results from Experiment 3: density plots for (a) mean deviation of participants' predictions from the model's prediction and (b) mean weight of advice. Numbers in each subplot indicate average values over all participants in the corresponding condition and error bars indicate one standard error.

4.2 Results

H7. **Deviation.** Contrary to the hypothesis, and in line with the findings from the first two experiments, there was no significant difference in participants' deviation from the model between CLEAR-2 and BB-8 ($t(798) = -0.87$, $p = 0.384$, see Figure 8a).

---

[6] We initially piloted a version of this study where participants were first shown an apartment and asked to predict its price, then shown the model and its prediction for that apartment, and finally asked to update their own prediction before moving on to the next apartment. We expected that the distribution of initial predictions for any given apartment would be the same regardless of which model participants were shown, but this turned out not be the case. Instead, we found that participants in the CLEAR-2 condition submitted initial predictions that were closer to the model's prediction—which they had not yet seen—compared to participants in other conditions ($t(239) = -3.42$, $p < 0.001$). We suspect this is because the CLEAR-2 condition is the easiest to simulate. Specifically, when making a prediction for a new apartment, participants in the CLEAR-2 condition could have internalized the model coefficients they saw in previous examples (i.e., $350,000 per bathroom and $1,000 per square foot) and used this to guide their initial prediction. This is an interesting potential benefit of transparent, simple models, but also poses a threat to validity for measuring the weight of advice given to different model presentations. To be able to compare weight of advice in different conditions, participants' initial predictions should not be influenced by the conditions they were assigned to, which is why we modified the design to collect initial predictions for all twelve apartments before participants were introduced to the model.

H8. **Weight of advice.** Weight of advice is not well defined when a participant's initial prediction matches the model's prediction (i.e., $u_1 = m$). For each condition, we therefore calculated the mean weight of advice over all participant–apartment pairs for which the participant's initial prediction did not match the model's prediction.[7] This can be viewed as calculating the mean conditional on there being a difference between the participant's and the model's predictions. Contrary to the hypothesis, and in line with the findings for deviation in the first two experiments, there was no significant difference in participants' weight of advice between the CLEAR-2 and BB-8 conditions ($t(819) = 1.27$, $p = 0.205$, see Figure 8b).

H9. **Humans vs. machines.** Contrary to the hypothesis, there was not a significant difference in participants' deviation from the model ($t(994) = 0.45$, $p = 0.655$) or in their weight of advice ($t(1005) = -0.38$, $p = 0.704$) between the BB-8 and EXPERT conditions. We expect that the difference between our results and those in [49] is due to participants getting more experience with the model (or expert) and its predictions over the course of twelve apartments in our experiment.

H10. **Detection of mistakes.** In contrast to the first two experiments, we did not find that participants in the clear conditions were less able to correct inaccurate predictions ($t(798) = -0.96$, $p = 0.337$ and $t(798) = -0.19$, $p = 0.847$ for the contrast of CLEAR-2 and CLEAR-8 with BB-2 and BB-8 for apartment 11 and apartment 12, respectively). We will investigate this further in the next experiment.

This experiment confirmed the findings from the first two experiments regarding deviation. That is, we again found that participants were no more likely to follow the predictions of a clear model with a small number of features than the predictions of a black-box model with a large number of features, this time as indicated by their weight of advice as well as by the deviation of their own predictions from the model's prediction. We also found that participants were no more or less likely to follow the predictions of a human expert than a black-box model. Finally, in contrast to the findings from the first two experiments, we did not find that participants in the clear conditions were less able to correct inaccurate predictions on unusual examples, a result that motivated our final experiment.

5 Experiment 4: Outlier focus and the detection of mistakes

Recall that, contrary to expectations, participants in the clear conditions in our first two experiments were less likely to notice when the model overpriced an apartment compared with black-box conditions (Figures 5a and 5b). In seeming contradiction, our third experiment revealed practically no difference across conditions in terms of participants' abilities to detect the model's sizable mistakes. In this section, we propose an explanation for these findings and support it with an additional experiment. The explanation rests on two conjectures. First, visual displays with a great amount of numerical information (e.g., model coefficients) might cause users not to notice peculiarities of a case under consideration due to information overload [1, 34, 36]. Second, when past predictions are visible, they might exert an anchoring effect [21, 68] on subsequent predictions, swaying new predictions in the direction of the past predictions.

In the first two experiments, participants made their final prediction while seeing their *simulation of the model's prediction* (Figure 2c). In contrast, participants in the third experiment made their final prediction while seeing their *own initial guess of the price* (Figure 7b). That is, there were different anchor values in the first and second experiments compared with the third.

---

[7]Between conditions, we found no significant difference in the fraction of times that participants' initial predictions matched the model's predictions.

Furthermore, within the first two experiments, anchor values differed by condition because they were influenced by the model presentation. Participants in the CLEAR-2 condition could simulate the model rather well relative to other conditions (Figures 3a and 4a). When the model overprices an apartment, simulating the model well might actually anchor the participant on a value that is too high. In addition, since clear models present more information, participants in these conditions might be unlikely to notice unusual apartment configurations. Together, these factors could explain their overly high final predictions.

In contrast, participants in the black-box conditions could not simulate the model easily, but, undistracted by model internals, might be more likely to notice unusual apartment configurations. Interestingly, participants in these conditions apparently (incorrectly) assumed the model would also take the unusual configurations into account and thus reported simulated values that were too low (relative to the model). When making their final predictions, participants in the black-box conditions therefore could have had two factors working in their favor: they were not overwhelmed by the model internals and they were anchored on the low values they had previously stated. Both of these factors could lower their final estimate for the unusual apartments and thus increase their predictive accuracy.

In brief, being overwhelmed by the internals of the clear models and possibly anchoring on prior estimates could together account for the observed differences in noticing unreasonable predictions in our first two experiments as opposed to our third experiment. To test this account, our fourth experiment was designed to remove possible sources of anchoring and then measure the influence of an "outlier focus" message that encouraged participants to take note of the configurations of the unusual apartments. Before running this experiment, we pre-registered three hypotheses:[8]

H11. **Effect of outlier focus.** Participants in the outlier focus conditions will exhibit different abilities to correct the model's mistakes compared to those in the control conditions (who are not shown outlier focus messages).
H12. **Effect of model transparency within control conditions.** Participants in the control conditions will exhibit different abilities to correct the model's mistakes based on whether they are shown a clear model or a black-box model.
H13. **Effect of model transparency within outlier focus conditions.** Participants in the outlier focus conditions will exhibit different abilities to correct the model's mistakes based on whether they are shown a clear model or a black-box model.

5.1 Experimental design

Similar to the first two experiments, we asked people to predict apartment prices with the help of a model. We considered four experimental conditions in a 2×2 design:
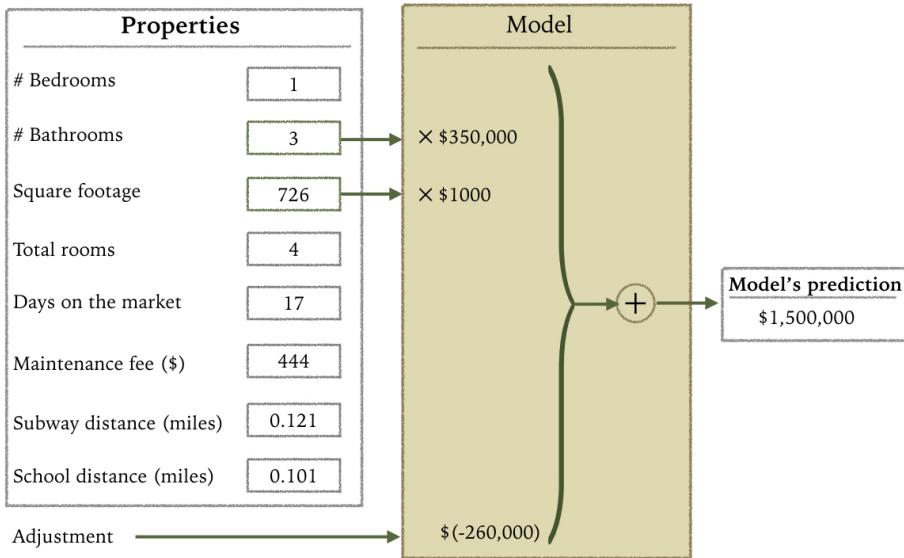
- Participants were randomly assigned to see the model internals (CLEAR), or see the model as a black box (BB).
- Participants were randomly assigned to see an explicit message concerning the unusual configurations of apartments (FOCUS), or not (NO-FOCUS).

Since there did not appear to be a strong effect of the number of features on people's abilities to detect the model's mistakes in our previous experiments, we only considered the 2-feature conditions. A screenshot of an apartment displayed for participants in the outlier focus condition is shown in Figure 9.

We again ran the experiment on Amazon Mechanical Turk. We excluded people who had participated in our first three

---

[8]Pre-registered hypotheses are available at `https://aspredicted.org/5xy8y.pdf`.

Figure 9: Apartment 6 in the outlier focus condition from Experiment 4. Participants were provided with an explicit outlier focus message (in red) when viewing apartment 6 and apartment 8 during the testing phase.

experiments, and recruited 800 new participants all of whom satisfied the selection criteria from our first three experiments. The participants were randomly assigned to the four conditions (CLEAR-FOCUS, $n = 202$; CLEAR-NO-FOCUS, $n = 195$; BB-FOCUS, $n = 201$; and BB-NO-FOCUS, $n = 202$) and each participant received a flat payment of $1.00.

We shortened the experiment by limiting the training phase and the first part of the testing phase to only five of the original ten apartments in each. We then added three more apartments to the end of the testing phase: two synthetically generated apartments with unusual configurations, with an apartment with a normal configuration (one-bedroom, one-bathroom, 788 square feet) in between. The first synthetically generated apartment was apartment 12 from our previous experiments (one-bedroom, three-bathroom, 726 square feet, referred to as *apartment 6* in this experiment). The second synthetically generated apartment had an even more unusual configuration (one-bedroom, three-bathroom, 350 square feet, referred to as *apartment 8* in this experiment), for which the model makes an unreasonably high prediction of the price. The order of the two synthetically generated apartments was randomized, while the apartment with the normal configuration (*apartment 7*) was always shown in the middle.

Participants went through the exact same training phase as in the first two experiments. For each apartment during the testing phase, participants were first shown the model's prediction and were asked to indicate how confident they were that the model was correct. Then they were asked to make their own prediction of the apartment's price and to indicate how confident they were in this prediction. To simplify the design and remove potential sources of anchoring, we did not ask participants to simulate the model's prediction in this experiment.
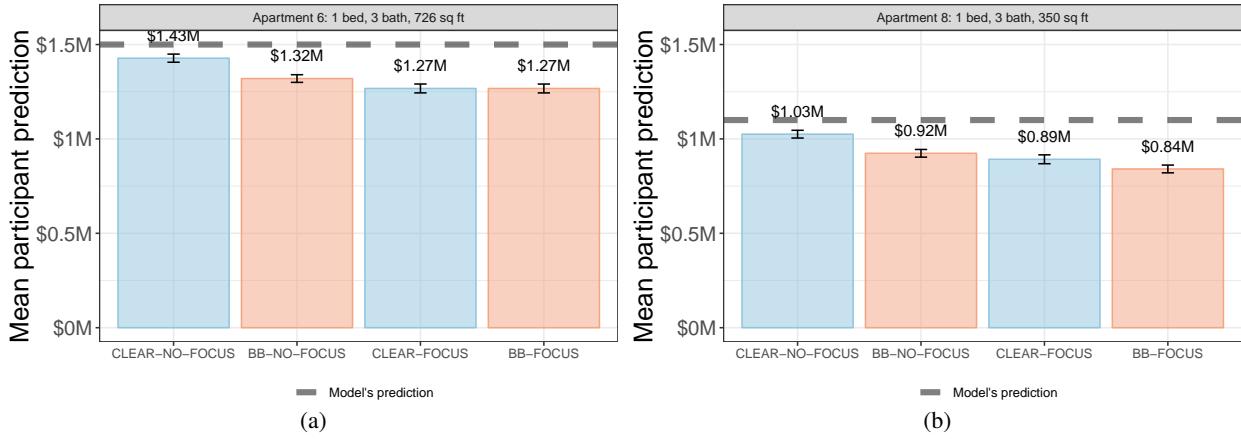
Figure 10: Results from Experiment 4: mean predictions of the prices of two unusual apartments for which the model should overestimate the price (a) apartment 6 and (b) apartment 8. Error bars indicate one standard error.

5.2 Results

Figure 10 shows the mean predictions of the prices of apartment 6 and apartment 8 from our fourth experiment. To test our hypotheses, we pre-registered the difference between a participant's prediction (u_a) and the model's prediction (m) as the dependent variable of interest: $u_a - m$.

H11. **Effect of outlier focus.** Participants in the outlier focus conditions differed from the model's prediction more, on average, than participants in the control conditions for both apartment 6 ($t(791) = -4.72, p < 0.001$) and apartment 8 ($t(795) = -5.00, p < 0.001$). That is, providing an outlier focus message improved participants' predictions.

H12. **Effect of model transparency within control conditions.** In the control conditions, predictions in the clear condition differed from the model's prediction less, on average, than predictions in the black-box condition for both apartment 6 ($t(393) = -3.65, p < 0.001$) and apartment 8 ($t(395) = -3.51, p < 0.001$). That is, in the absence of an outlier focus message, participants in the clear condition again performed worse by deviating less from the model's unreasonably high predictions.

H13. **Effect of model transparency within outlier focus conditions**. With an outlier focus message, we found no significant effect of model transparency on how much participants' predictions differed from the model's predictions ($t(401) = -0.004, p = 0.996$ for apartment 6 and $t(394) = -1.64, p = 0.101$ for apartment 8). This result is consistent with the idea that an outlier focus message can help participants react to the peculiarities that get obscured by the overwhelming information presented by clear models.

Taken together, these results suggest a simple explanation consistent with our information overload conjecture for why participants who were shown clear models performed poorly on apartments with unusual configurations. Without an outlier focus message, exposing the details of a model's internals can cause people to miss unusual inputs to the model. With such a message, however, this effect disappears. This both highlights an unintended downside of clear models and offers a simple intervention to neutralize the effect.

6 Discussion and future work

Our experiments yielded several surprising results. On typical examples, we saw no significant difference between a transparent model with few features and a black-box model with many features in terms of how closely participants followed the model's predictions. We also saw that people would have been better off simply following the models rather than adjusting their predictions. Even more surprisingly, we found that transparent models had the unwanted effect of impairing people's ability to correct inaccurate predictions, seemingly due to people being overwhelmed by the additional information that the transparent model presented. When we tested an intervention to counter information overload, this unwanted effect was mitigated, consistent with the idea that the information provided in a transparent model inhibited people's ability to notice unusual inputs. Several empirical findings from post-hoc analyses were also consistent with the idea that presenting too much information can be detrimental. In Experiments 1 and 2, the decision aid that presented participants with the *most* information—the clear, eight-feature model—led to the *worst* performance on simulation, deviation, and prediction of all the models we tested.

These findings suggest new ways to present models to decision makers. When technically possible, it may be helpful for decision aids to alert users that the case under consideration may be an outlier. This could be achieved by training an auxiliary model to detect such cases. In addition, it may be prudent to have practitioners provide their own predictions before seeing the details or predictions of a model. Doing so could encourage decision makers to inspect the details of each case, increasing the likelihood that they notice any unusual inputs. Lastly, despite potential benefits of transparent models, it may be detrimental to expose model internals by default, as doing so might cause people to miss important details. Instead, model internals could be hidden until the user requests to see them. Testing these suggestions empirically would be a natural direction for future research.

None of this is to say that transparency and parsimony should be ignored or avoided when designing decision aids. Instead, our results underscore the point that there are many possible goals in designing interpretable models, and that one should rely on rigorous experimentation over intuition to assess whether those goals are met.

A limitation of our work is that the experiments focused on just one model, in one domain, for a subset of scenarios in which interpretability might matter. Future extensions to other tasks (e.g., classification), models (e.g, decision trees or rule lists), users (e.g., data scientists or domain experts), and domains (e.g., medicine, hiring, judicial decisions) may yield different results than those found here. Even within the design of our experiments, although we have controlled for many potential confounds, we cannot rule out that there could have been specifics about the particular models we used beyond simply the number of features and transparency that influenced our results. For instance, people might have found the combination of features used in the two-feature model (i.e., bathrooms and square feet) to be less intuitive than other possible two-feature combinations (e.g., bedrooms and bathrooms) or a three-feature model (e.g., bedrooms, bathrooms, and square feet). Also, those who saw a two-feature model were in the unique position of having access to more features than the model did. For this reason, people might have thought a model that uses two out of eight features was overly simplistic. Perhaps if they were told that incorporating the remaining six features would not improve the model's performance, they would have viewed the two-feature model differently. Conversely, it could have been the case that specifics of the eight-feature model led people to question its quality. For example, the negative weight this model placed on the total number of rooms (to account for correlations with the number of bedrooms and bathrooms) might have been confusing or mistakenly viewed as wrong, leading people to follow it less closely than they would have otherwise. Although these factors make it difficult to isolate exactly *why* we do not see differences in deviation or prediction quality across these different models, it still remains the case that using the intuitively simpler and more transparent model did not produce higher quality decisions than using the more complex, opaque one and

even hampered people's ability to correct models' inaccurate predictions.

At the same time, there is still a long list of reasons why interpretability might be desirable. First, in some domains, interpretability may play an important role in people's willingness to adopt a model on ethical grounds. For instance, if a model is used to make judicial decisions, policy makers may demand transparency to be assured that these decisions are not being made on the basis of disallowed criteria like race or proxies for race. Second, access to model internals allows people to better anticipate the inputs for which a model might go wrong. In fact, we leveraged this aspect of our own linear regression model for apartment prices to design the unusual apartment configurations used in our experiments, since we could easily see that it would place unreasonably high value on additional bathrooms while keeping other features fixed; detailed debugging and analysis would have been more difficult with a black-box model. Third, in settings where it is desirable to have a model that is easy to simulate (e.g., decision rules used by field workers), our experiments suggest that simple, transparent models had an advantage: the clear, two-feature model won out on being easiest to simulate. Fourth, while we did not investigate adoption of decision aids, it might be the case that people are more likely to adopt simple models over complex ones because they find them more appealing [35]. Given that simple, transparent models improved predictions as effectively as any other model over the baseline of not having a model at all, if it is the case that people are more willing to adopt them, there could be substantial benefits in terms of decision quality. The results of our experiments should be considered jointly with these observations, some of which are quite encouraging for the use of simple, transparent models as decision aids.

In light of recent advances, it is likely that in the future more and more decisions will be made with the aid of machine learning. As we transition to this future, it is also likely there will be a demand for models that people can interpret. We hope that the often surprising results in this paper will reinforce that the behavior of decision makers, and not only the intuitions of modelers, should guide the creation of interpretable decisions aids.

References

[1] Russell L Ackoff. Management misinformation systems. *Management Science*, 14(4):B–141–B–274, 1967. URL `https://doi.org/10.1287/mnsc.14.4.B147`.

[2] Oscar Alvarado and Annika Waern. Towards algorithmic experience: Initial efforts for social media contexts. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 286. ACM, 2018. URL `http://doi.acm.org/10.1145/3173574.3173860`.

[3] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica*, 2016.

[4] Thomas Åstebro and Samir Elhedhli. The effectiveness of simple decision heuristics: Forecasting commercial success for early-stage ventures. *Management Science*, 52(3):395–409, 2006. URL `https://doi.org/10.1287/mnsc.1050.0468`.

[5] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015.

[6] Max H Bazerman. Norms of distributive justice in interest arbitration. *ILR Review*, 38(4):558–570, 1985.

[7] James Bennett, Stan Lanning, et al. The netflix prize. In *Proceedings of KDD Cup and Workshop*, 2007.

[8] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 'it's reducing a human being to a percentage': Perceptions of justice in algorithmic decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 377. ACM, 2018. URL `http://doi.acm.org/10.1145/3173574.3173951`.

[9] Robert C Blattberg and Stephen J Hoch. Database models and managerial intuition: 50% model+ 50% manager. *Management Science*, 36(8):887–899, 1990. URL `https://doi.org/10.1287/mnsc.36.8.887`.

[10] Taina Bucher. The algorithmic imaginary: exploring the ordinary affects of facebook algorithms. *Information, Communication & Society*, 20(1):30–44, 2017.

[11] Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on psychological science*, 6(1):3–5, 2011.

[12] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2015.

[13] Krista Casler, Lydia Bickel, and Elizabeth Hackett. Separate but equal? a comparison of participants and data gathered via amazon's mturk, social media, and face-to-face behavioral testing. *Computers in Human Behavior*, 29(6):2156–2160, 2013.

[14] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[15] Eric Colson. Using human and machine processing in recommendation systems. In *First AAAI Conference on Human Computation and Crowdsourcing*, 2013.

[16] Alexander Coppock. Generalizing from survey experiments conducted on mechanical turk: A replication approach. *Political Science Research and Methods*, 7(3):613–628, 2019.

[17] Jeffrey Dastin. Amazon scraps secret AI recruiting tool that showed bias against women. 2018.

[18] Robyn M Dawes. The robust beauty of improper linear models in decision making. *American psychologist*, 34(7): 571, 1979.

[19] Robyn M Dawes, David Faust, and Paul E Meehl. Clinical versus actuarial judgment. *Science*, 243(4899): 1668–1674, 1989.

[20] Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114–126, 2015.

[21] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, 64(3):1155–1170, 2016. URL https://doi.org/10.1287/mnsc.2016.2643.

[22] Jaap J Dijkstra. User agreement with incorrect expert system advice. *Behaviour & Information Technology*, 18(6): 399–411, 1999.

[23] Jaap J Dijkstra, Wim BG Liebrand, and Ellen Timminga. Persuasiveness of expert systems. *Behaviour & Information Technology*, 17(3):155–163, 1998.

[24] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[25] Mary T Dzindolet, Linda G Pierce, Hall P Beck, and Lloyd A Dawe. The perceived utility of human and automated aids in a visual detection task. *Human Factors*, 44(1):79–94, 2002.

[26] Motahhare Eslami, Sneha R Krishna Kumaran, Christian Sandvig, and Karrie Karahalios. Communicating algorithmic process in online behavioral advertising. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 432. ACM, 2018. URL http://doi.acm.org/10.1145/3173574.3174006.

[27] Vivian Giang. The potential hidden bias in automated hiring systems. Accessed at https://www.fastcompany.com/40566971/the-potential-hidden-bias-in-automated-hiring-systems/, 2018.

[28] Gerd Gigerenzer and Daniel G Goldstein. Reasoning the fast and frugal way: models of bounded rationality. *Psychological review*, 103(4):650, 1996.

[29] Francesca Gino and Don A. Moore. Effects of task difficulty on use of advice. *Journal of Behavioral Decision Making*, 20(1):21–35, 2007.

[30] Alyssa Glass, Deborah L McGuinness, and Michael Wolverton. Toward establishing trust in adaptive agents. In *Proceedings of the 13th International Conference on Intelligent User Interfaces (IUI)*, 2008.

[31] William M Grove, David H Zald, Boyd S Lebow, Beth E Snitz, and Chad Nelson. Clinical versus mechanical prediction: a meta-analysis. *Psychological assessment*, 12(1):19, 2000.

[32] Todd M Gureckis, Jay Martin, John McDonnell, Alexander S Rich, Doug Markant, Anna Coenen, David Halpern, Jessica B Hamrick, and Patricia Chan. psiTurk: An open-source framework for conducting replicable behavioral experiments online. *Behavior Research Methods*, 48(3):829–842, 2016.

[33] David J Hand and William E Henley. Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 160(3):523–541, 1997.

[34] Jacob Jacoby. Perspectives on information overload. *Journal of consumer research*, 10(4):432–435, 1984.

[35] Jongbin Jung, Connor Concannon, Ravi Shro, Sharad Goel, and Daniel G. Goldstein. Simple rules for complex decisions. *arXiv preprint arXiv:1702.04690*, 2017.

[36] Kevin Lane Keller and Richard Staelin. Effects of quality and quantity of information on decision effectiveness. *Journal of consumer research*, 14(2):200–213, 1987.

[37] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human decisions and machine predictions. *The quarterly journal of economics*, 133(1):237–293, 2017.

[38] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

[39] Igor Kononenko. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in medicine*, 23(1):89–109, 2001.

[40] Josua Krause, Aritra Dasgupta, Jordan Swartz, Yindalon Aphinyanaphongs, and Enrico Bertini. A workflow for visual diagnostics of binary classifiers using instance-level explanations. In *Proceedings of IEEE Conference and Visual Analytics Science and Technology*, 2017.

[41] Isaac Lage, Andrew Slavin Ross, Been Kim, Samuel J. Gershman, and Finale Doshi-Velez. Human-in-the-loop interpretability prior. In *Advances in Neural Information Processing Systems*, 2018.

[42] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. Interpretable & explorable approximations of black box models. In *FATML Workshop*, 2017.

[43] Cynthia CS Liem, Markus Langer, Andrew Demetriou, Annemarie MF Hiemstra, Achmadnoer Sukma Wicaksana, Marise Ph Born, and Cornelius J König. Psychology meets machine learning: Interdisciplinary perspectives on algorithmic job candidate screening. In *Explainable and Interpretable Models in Computer Vision and Machine Learning*, pages 197–253. Springer, 2018.

[44] Brian Y Lim and Anind K Dey. Design of an intelligible mobile context-aware application. In *Proceedings of the 13th international conference on human computer interaction with mobile devices and services*, pages 157–166. ACM, 2011. URL http://doi.acm.org/10.1145/2037373.2037399.

[45] Brian Y Lim, Anind K Dey, and Daniel Avrahami. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the 2009 CHI Conference on Human Factors in Computing Systems*, pages 2119–2128. ACM, 2009. URL http://doi.acm.org/10.1145/1518701.1519023.

[46] Zachary C Lipton. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*, 2016.

[47] John DC Little. Models and managers: The concept of a decision calculus. *Management Science*, 16(8):B–466, 1970. URL https://doi.org/10.1287/mnsc.16.8.B466.

[48] Jia Liu and Olivier Toubia. A semantic approach for estimating consumer content preferences from online search queries. *Marketing Science*, 37(6):855–1052, 2018. URL `https://doi.org/10.1287/mksc.2018.1112`.

[49] Jennifer M. Logg. Theory of machine: When do people rely on algorithms? Harvard Business School NOM Unit Working Paper No. 17-086, 2017.

[50] Yin Lou, Rich Caruana, and Johannes Gehrke. Intelligible models for classification and regression. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2012.

[51] Yin Lou, Rich Caruana, Johannes Gehrke, and Giles Hooker. Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2013.

[52] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30 (NIPS)*, 2017.

[53] Winter Mason and Siddharth Suri. Conducting behavioral research on amazon's mechanical turk. *Behavior research methods*, 44(1):1–23, 2012.

[54] Rishabh Mehrotra, James McInerney, Hugues Bouchard, Mounia Lalmas, and Fernando Diaz. Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness & satisfaction in recommendation systems. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 2243–2251. ACM, 2018. URL `http://doi.acm.org/10.1145/3269206.3272027`.

[55] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 2018.

[56] Tim Miller, Piers Howe, and Liz Sonenberg. Explainable AI: Beware of inmates running the asylum. *arXiv preprint arXiv:1712.00547*, 2017.

[57] Safiya Umoja Noble. *Algorithms of Oppression: How search engines reinforce racism*. NYU Press, 2018.

[58] Dilek Önkal, Paul Goodwin, Mary Thomson, and Sinan Gönül. The relative influence of advice from human experts and statistical methods on forecast adjustments. *Journal of Behavioral Decision Making*, 22:390–409, 2009.

[59] Gabriele Paolacci and Jesse Chandler. Inside the turk: Understanding mechanical turk as a participant pool. *Current Directions in Psychological Science*, 23(3):184–188, 2014.

[60] Marianne Promberger and Jonathan Baron. Do patients trust computers? *Journal of Behavioral Decision Making*, 19(5):455–468, 2006.

[61] Emilee Rader and Rebecca Gray. Understanding user beliefs about algorithmic curation in the facebook news feed. In *Proceedings of the 2015 CHI Conference on Human Factors in Computing Systems*, pages 173–182. ACM, 2015. URL `http://doi.acm.org/10.1145/2702123.2702174`.

[62] Emilee Rader, Kelley Cotter, and Janghee Cho. Explanations as mechanisms for supporting algorithmic transparency. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 103. ACM, 2018. URL `http://doi.acm.org/10.1145/3173574.3173677`.

[63] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 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 (KDD)*, 2016.

[64] Christian Rudder. *Dataclysm: Love, Sex, Race, and Identity–What Our Online Lives Tell Us about Our Offline Selves*. Crown, 2014.

[65] Cynthia Rudin and Berk Ustun. Optimized scoring systems: Toward trust in machine learning for healthcare and criminal justice. *Applied Analytics*, 48(5):449–466, 2018. URL `https://doi.org/10.1287/inte.2018.0957`.

[66] Paul JH Schoemaker and C Carter Waid. An experimental comparison of different approaches to determining weights in additive utility models. *Management Science*, 28(2):182–196, 1982. URL `https://doi.org/10.1287/mnsc.28.2.182`.

[67] Richard Tomsett, David Braines, Dan Harborne, Alun Preece, and Supriyo Chakraborty. Interpretable to whom? A role-based model for analyzing interpretable machine learning systems. In *2018 Workshop on Human Interpretability in Machine Learning*, 2018.

[68] Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. *science*, 185(4157): 1124–1131, 1974.

[69] Berk Ustun and Cynthia Rudin. Supersparse linear integer models for optimized medical scoring systems. *Machine Learning Journal*, 102(3):349–391, 2016.

[70] Sara Wachter-Boettcher. Why you can't trust ai to make unbiased hiring decisions. Accessed at `http://time.com/4993431/ai-recruiting-tools-do-not-/`, 2017.

[71] Martin Wattenberg, Fernanda Viégas, and Moritz Hardt. Attacking discrimination with smarter machine learning. Accessed at `https://research.google.com/bigpicture/attacking-discrimination-in-ml/`, 2016.

[72] Ilan Yaniv. Receiving other people's advice: Influence and benefit. *Organizational Behavior and Human Decision Processes*, 93:1–13, 2004.

[73] Mike Yeomans, Anuj Shah, Sendhil Mullainathan, and Jon Kleinberg. Making sense of recommendations. *Preprint at http://scholar. harvard. edu/files/sendhil/files/recommenders55_01. pdf*, 2016.

.1 Appendix: Instructions from the CLEAR-2 Condition in Experiment 1

The following instructions were shown to participants assigned to the CLEAR-2 condition in our first experiment on Mechanical Turk. The instructions for other conditions and experiments were adapted from these instructions with minimal changes.

Instructions

!! IMPORTANT !! Your session will expire in 60 minutes. Please make sure to complete the HIT in 60 minutes!

- You are here to predict **New York City apartment prices in the <u>Upper West Side</u>** with the help of a model.
- There will be a <u>training phase</u> and a <u>testing phase</u>:
    - In the training phase, you will see examples of apartments along with what the model predicted they sold for and the actual price they sold for.
    - In the testing phase, you will see new apartments and make your own prediction about what the model will predict and what the actual price is.

Next ➜

Instructions

- You will see these properties for each apartment:
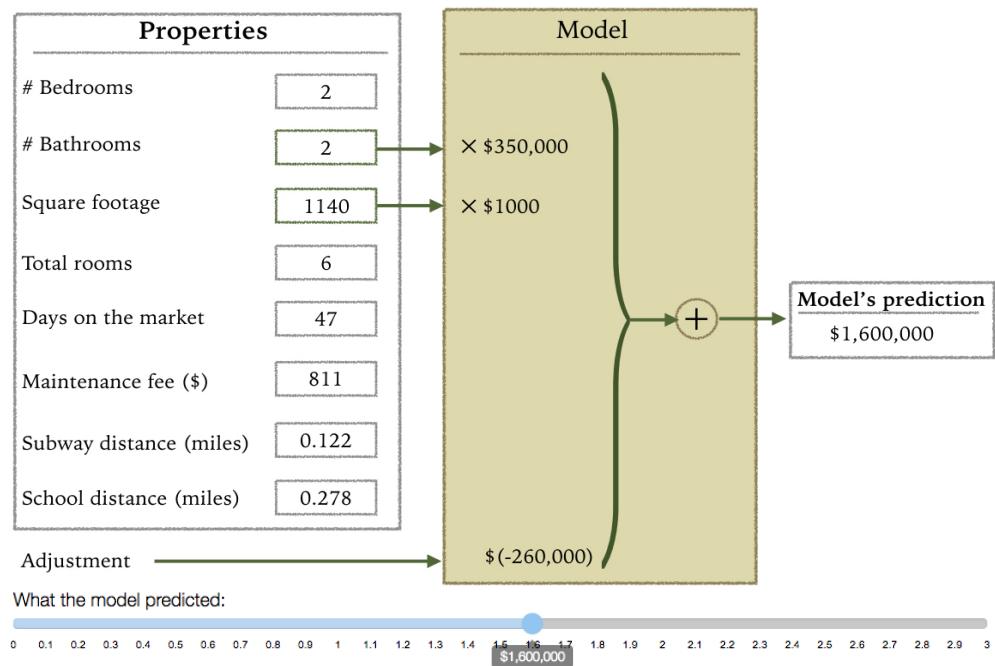
Properties	
# Bedrooms	2
# Bathrooms	2
Square footage	1140
Total rooms	6
Days on the market	47
Maintenance fee ($)	811
Subway distance (miles)	0.122
School distance (miles)	0.278

Instructions

- A model predicts apartment prices. We will explain how this model works in the next page.
  - This model uses # Bathrooms and Square footage of the apartment to make its prediction.
  - The graph at the bottom shows this price visually.

Properties	
# Bedrooms	2
# Bathrooms	2
Square footage	1140
Total rooms	6
Days on the market	47
Maintenance fee ($)	811
Subway distance (miles)	0.122
School distance (miles)	0.278

Model

× $350,000

× $1000

+

Model's prediction
$1,600,000

Adjustment ⟶ $ (-260,000)

What the model predicted:

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2 2.1 2.2 2.3 2.4 2.5 2.6 2.7 2.8 2.9 3
$1,600,000
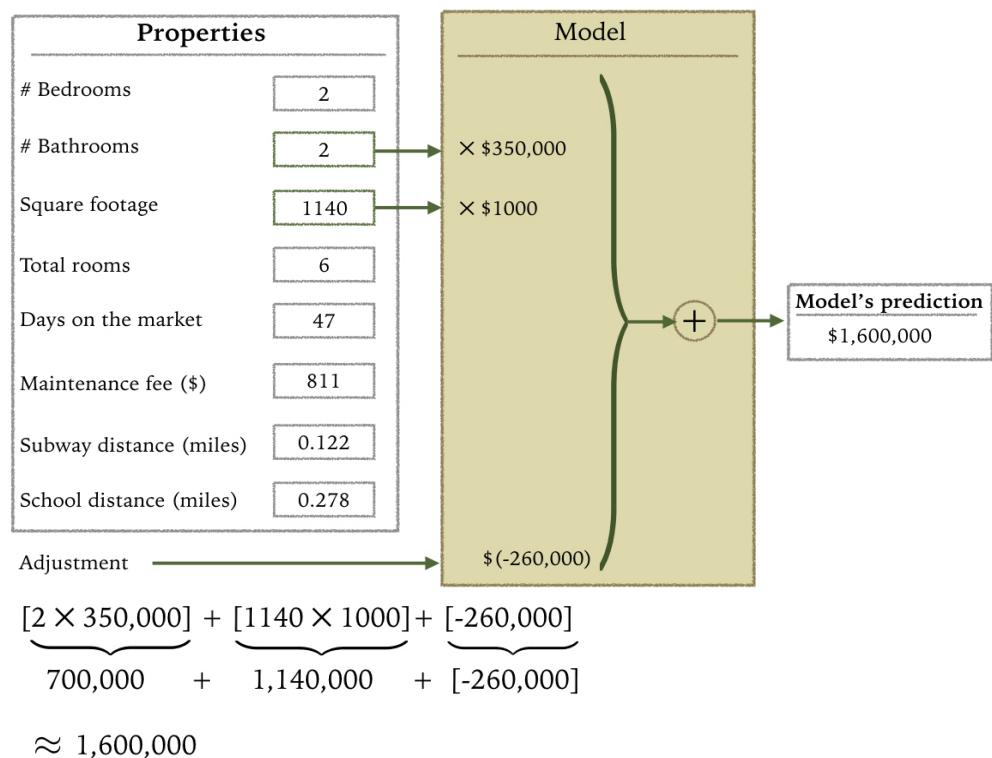
← Previous

Next →

Instructions

- Here is how the model has made its prediction:
- Each bathroom is worth $350,000. Therefore, $350,000 is multiplied by the number of bathrooms and added to the price. This is repeated for Square footage. Finally, **the adjustment factor of $260,000 is subtracted** and a price is predicted.

Properties		Model
# Bedrooms	2	
# Bathrooms	2	✕ $350,000
Square footage	1140	✕ $1000
Total rooms	6	
Days on the market	47	**Model's prediction** $1,600,000
Maintenance fee ($)	811	
Subway distance (miles)	0.122	
School distance (miles)	0.278	
Adjustment		$(-260,000)

$$[2 \times 350{,}000] + [1140 \times 1000] + [-260{,}000]$$

$$700{,}000 \quad + \quad 1{,}140{,}000 \quad + \quad [-260{,}000]$$

$$\approx 1{,}600{,}000$$

← Previous   Next →

Training Phase Instructions
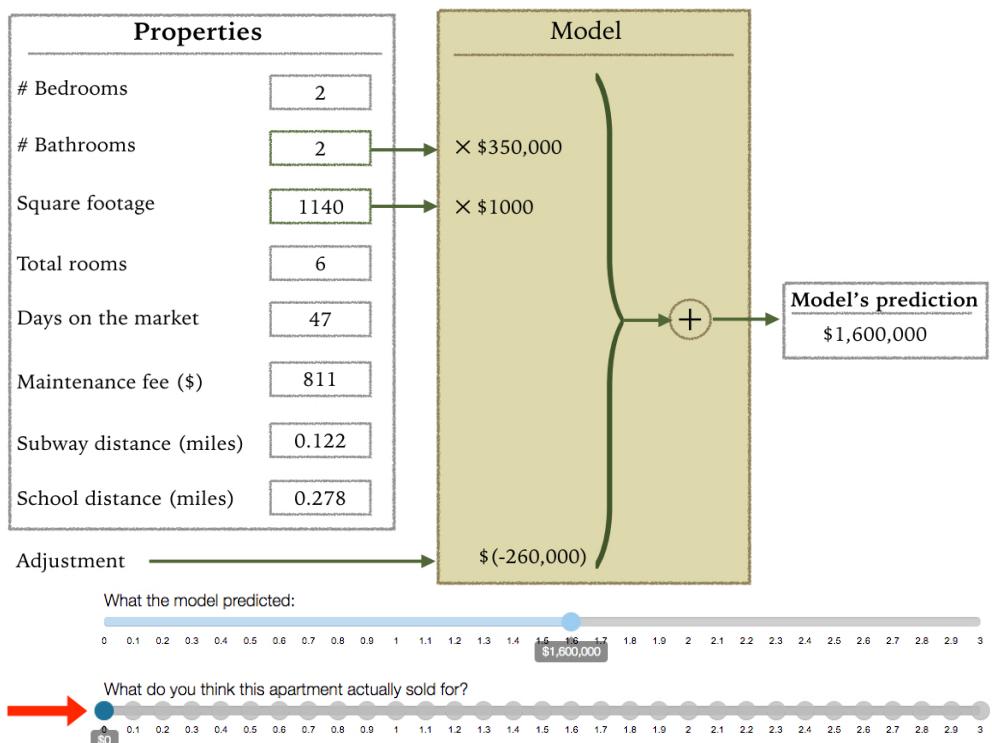
- There will be ten apartments in the training phase.
- For each apartment, you will complete the following two steps:

Training Phase Instructions-Step 1

- In step 1, given the model's prediction, you will state what you think the apartment actually sold for:

Properties	
# Bedrooms	2
# Bathrooms	2
Square footage	1140
Total rooms	6
Days on the market	47
Maintenance fee ($)	811
Subway distance (miles)	0.122
School distance (miles)	0.278

Model

× $350,000

× $1000

+

Model's prediction
$1,600,000

Adjustment → $ (-260,000)

What the model predicted:

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2 2.1 2.2 2.3 2.4 2.5 2.6 2.7 2.8 2.9 3
$1,600,000

What do you think this apartment actually sold for?

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2 2.1 2.2 2.3 2.4 2.5 2.6 2.7 2.8 2.9 3
$0

[← Previous]   [Next →]

Training Phase Instructions-Step 2

- In step 2, you will see what this apartment actually sold for and how you and the model did:
- There are three graphs at the bottom:
    - The first graph shows the model's prediction of the price of this apartment.
    - The second graph shows what you thought this apartment actually sold for.
    - The third graph shows what this apartment actually sold for.

Properties		Model	
# Bedrooms	2		
# Bathrooms	2	✕ $350,000	
Square footage	1140	✕ $1000	
Total rooms	6		
Days on the market	47		**Model's prediction** $1,600,000
Maintenance fee ($)	811	(+)	
Subway distance (miles)	0.122		**Actual price** $1,800,000
School distance (miles)	0.278		

Adjustment ———————————————→ $ (-260,000)

What the model predicted:

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2 2.1 2.2 2.3 2.4 2.5 2.6 2.7 2.8 2.9 3
$1,600,000

The price you chose in step 1.

What you thought this apartment actually sold for:

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2 2.1 2.2 2.3 2.4 2.5 2.6 2.7 2.8 2.9 3
$0

What this apartment actually sold for:

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2 2.1 2.2 2.3 2.4 2.5 2.6 2.7 2.8 2.9 3
$1,800,000

← Previous

Next →

Instructions

- Once you have reviewed all ten apartments in the <u>training phase</u>, you will move to the <u>testing phase</u>.
- In the testing phase, you will see twelve new apartments and you will guess the price each was sold for.
- **NOTE**: You will not see the actual prices for these apartments in the testing phase. Once you are done, you will see how you did overall.
- For each apartment, you will complete the following three steps:

Testing Phase Instructions-Step 1

- In step 1, you will state what you think the model will predict and how confident you are that the model will make that prediction:

Properties		Model
# Bedrooms	1	
# Bathrooms	1	✕ $350,000
Square footage	841	✕ $1000
Total rooms	3.5	
Days on the market	36	⊕
Maintenance fee ($)	505	
Subway distance (miles)	0.122	
School distance (miles)	0.278	
Adjustment		$ (-260,000)

What do you think the model will predict?

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2 2.1 2.2 2.3 2.4 2.5 2.6 2.7 2.8 2.9 3

$800,000

How confident are you the model will predict this?

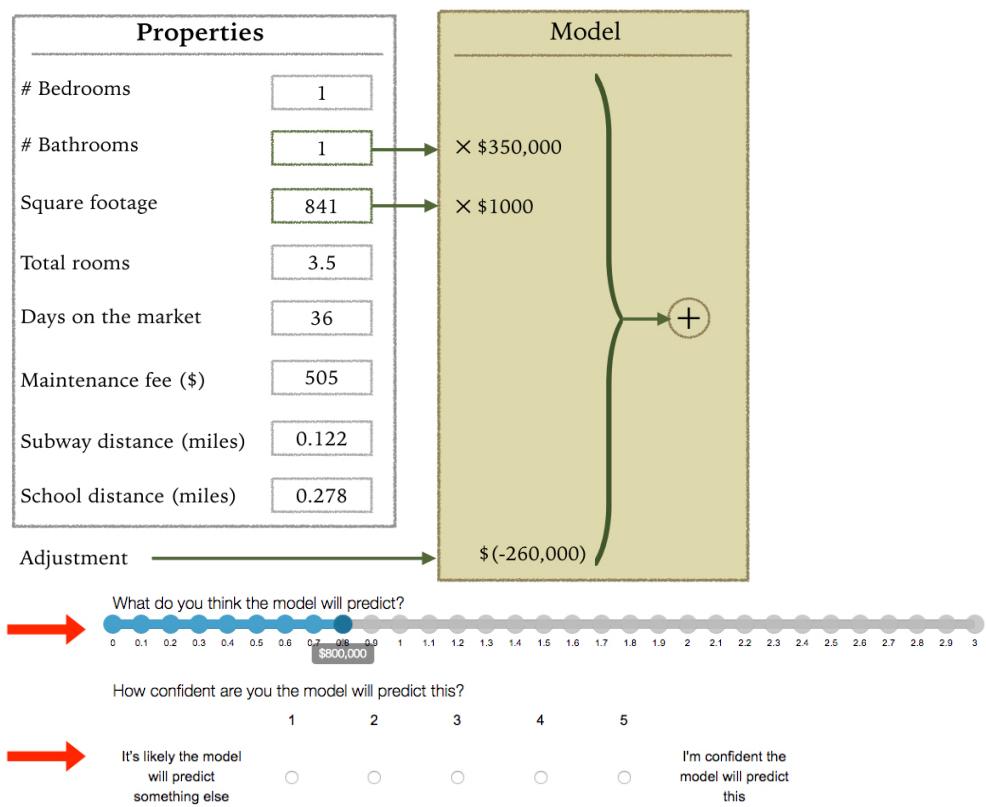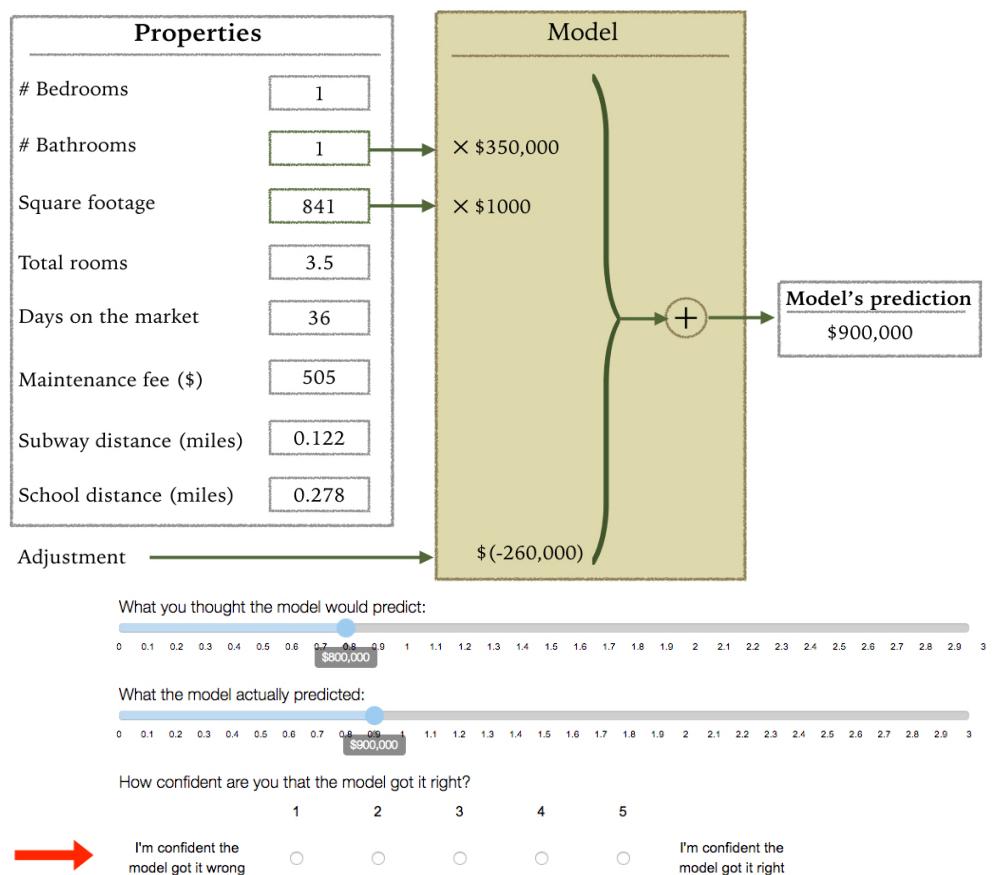|	1	2	3	4	5	
It's likely the model will predict something else	○	○	○	○	○	I'm confident the model will predict this

← Previous   Next ➡

Testing Phase Instructions-Step 2

- In step 2, you will see what the model predicts and you will state how confident you are that the model made the right prediction:

Properties

# Bedrooms	1
# Bathrooms	1
Square footage	841
Total rooms	3.5
Days on the market	36
Maintenance fee ($)	505
Subway distance (miles)	0.122
School distance (miles)	0.278

Model

× $350,000

× $1000

+

Adjustment $(-260,000)$

Model's prediction
$900,000

What you thought the model would predict:

0   0.1   0.2   0.3   0.4   0.5   0.6   0.7   0.8   0.9   1   1.1   1.2   1.3   1.4   1.5   1.6   1.7   1.8   1.9   2   2.1   2.2   2.3   2.4   2.5   2.6   2.7   2.8   2.9   3

$800,000

What the model actually predicted:

0   0.1   0.2   0.3   0.4   0.5   0.6   0.7   0.8   0.9   1   1.1   1.2   1.3   1.4   1.5   1.6   1.7   1.8   1.9   2   2.1   2.2   2.3   2.4   2.5   2.6   2.7   2.8   2.9   3

$900,000

How confident are you that the model got it right?

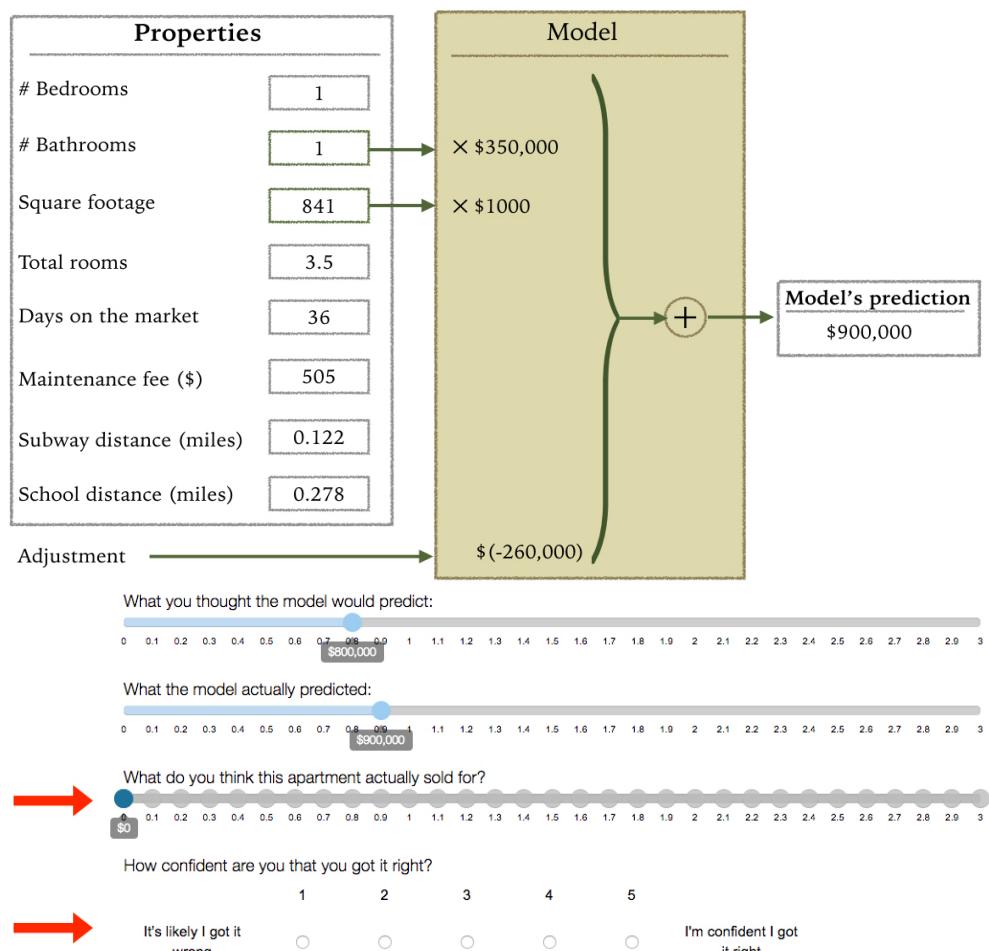|	1	2	3	4	5	
I'm confident the model got it wrong	○	○	○	○	○	I'm confident the model got it right

← Previous

Next →

Testing Phase Instructions-Step 3

- In step 3, given the model's prediction, you will state what you think this apartment actually sold for and your confidence:

Properties	
# Bedrooms	1
# Bathrooms	1
Square footage	841
Total rooms	3.5
Days on the market	36
Maintenance fee ($)	505
Subway distance (miles)	0.122
School distance (miles)	0.278

Model

× $350,000

× $1000

+

Model's prediction
$900,000

Adjustment $ (-260,000)

What you thought the model would predict:

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2 2.1 2.2 2.3 2.4 2.5 2.6 2.7 2.8 2.9 3
$800,000

What the model actually predicted:

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2 2.1 2.2 2.3 2.4 2.5 2.6 2.7 2.8 2.9 3
$900,000

What do you think this apartment actually sold for?

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2 2.1 2.2 2.3 2.4 2.5 2.6 2.7 2.8 2.9 3
$0

How confident are you that you got it right?

	1	2	3	4	5	
It's likely I got it wrong	○	○	○	○	○	I'm confident I got it right

← Previous   Next →

Instructions

- Once you are done with twelve apartments in the testing phase, you will see your results and how you did overall.
- You are now done with all the instructions. Thanks for participating in this experiment!

You must provide correct answer to the following question to proceed:

Each apartment has the following 8 properties:

Bedrooms, # Bathrooms, Square footage, Total rooms, Days on the market, Maintenance fee, Subway distance, and School distance

How many of these apartment properties does the model use to make its prediction?

- ○ 1 property
- ○ 2 properties
- ○ 3 properties
- ○ 4 properties
- ○ 5 properties
- ○ 6 properties
- ○ 7 properties
- ○ 8 properties

Submit

← Previous   Next →

Starting the Training Phase...

- You will now start the training phase.
- You will see ten apartments, the price that the model predicted, and the price that they were actually sold for.
- Pay attention to apartment properties and how they relate to the actual price and the model's prediction.
- **You won't be able to go back to the training phase once you get to the testing phase!**

← Previous

Begin training phase ➡

.2 Appendix: Apartment Selection Details

We used the following procedure to construct a set of ten apartments which are representative in terms of the models' prediction errors $(m-a)$. First we selected all apartments for which the rounded predictions of the two- and eight-feature models agreed. Then we randomly sampled 5,000 sets of ten such apartments, computed the errors the model made on each apartment, and sorted them within each set to obtain the largest error, second largest error, and so on. We then computed the average largest error across all 5,000 sets and rounded it to the nearest $100K. We repeated this for the second through tenth largest errors. This resulted in the following ten average error values: -\$500K, -\$300K, -\$200K, -\$200K, -\$100K, \$0, \$0, \$100K, \$100K, \$300K.

For each of these ten error values, we randomly selected two apartments (for which the difference between the rounded model prediction and the rounded actual price matched the error value) and we randomly assigned one to the training and one to the testing phase. To ensure participants would see a good variety of apartment configurations—defined as the combination of number of bedrooms and number of bathrooms—this process was repeated until neither the training nor the testing set contained more than three apartments with the same configuration. Tables 1 and 3 show the configurations of apartments that were used during the training and testing phase of each of our experiments, respectively. Tables 2 and 4 show the predictions and errors of the two- and eight-feature models on each of the apartments.

Apartment ID	Apartment configurations							
	Bedrooms	Bathrooms	Square footage	Total rooms	Days on the market	Maintenance fee	Distance from the subway (miles)	Distance from a school (miles)
1	1	1	750	3	51	947	0.179	0.104
2	1	1	550	3	90	409	0.122	0.278
3	2	1	800	4	36	1160	0.218	0.365
4	2	1	850	4	30	1720	0.105	0.153
5	1	1	550	3	135	442	0.231	0.124
6	0	1	540	2.5	72	332	0.064	0.271
7	3	2	1990	6	213	1280	0.183	0.329
8	2	1	1150	4	37	1500	0.129	0.351
9	0	1	540	2.5	59	331	0.064	0.271
10	2	2	1300	5	39	1110	0.110	0.250

Table 1: Configuration of the apartments used in experiments 1, 2, and 3 during the training phase. In Experiment 4, apartments 4, 5, 6, 8, and 10 were used.

Apartment ID	two-feature model			eight-feature model		
	prediction	error	error fraction	prediction	error	error fraction
1	840,000	-9,000	0.011	768,930	-80,070	0.094
2	640,000	-10,000	0.015	632,010	-17,990	0.028
3	890,000	241,000	0.371	893,500	244,500	0.377
4	940,000	115,000	0.139	850,600	25,600	0.031
5	640,000	-184,000	0.223	614,880	-209,120	0.254
6	630,000	175,000	0.385	550,080	95,080	0.209
7	2,430,000	-470,000	0.162	2,417,800	-482,200	0.166
8	1,240,000	90,000	0.078	1,195,600	45,600	0.04
9	630,000	-165,000	0.208	552,790	-242,210	0.305
10	1,740,000	-260,000	0.13	1,701,100	-298,900	0.149

Table 2: Prediction, prediction error (i.e., $m - a$), and prediction error fraction (i.e., $(m - a)/a$) of the two- and eight-feature models on the apartments used in the training phase of our experiments.

Apartment ID	Apartment configurations							
	Bedrooms	Bathrooms	Square footage	Total rooms	Days on the market	Maintenance fee	Distance from the subway (miles)	Distance from a school (miles)
1	1	1	925	3	80	954	0.173	0.312
2	2	1	1080	5	39	846	0.207	0.212
3	3	2	1530	5	15	1550	0.226	0.251
4	2	2	1140	4.5	93	863	0.122	0.278
5	1	1	540	3	11	437	0.202	0.199
6	0	1	540	2.5	74	341	0.122	0.278
7	2	1	1240	4.5	32	1370	0.081	0.262
8	2	2	1240	4.5	14	906	0.178	0.225
9	2	1	1250	5	23	1480	0.089	0.281
10	1	1	532	2.5	20	388	0.122	0.278
11	1	2	750	3	225	825	0.159	0.144
12	1	3	726	4	17	444	0.121	0.101
13	1	1	788	3.5	51	473	0.122	0.278
14	1	3	350	4	13	430	0.221	0.131

Table 3: Configuration of the apartments used in our experiments during the testing phase. Apartments 1–12 were used in experiments 1, 2, and 3. In Experiment 4, apartments 1, 6, 8, 9, 10, 12 ("Apartment 6" in Experiment 4), 13 ("Apartment 7" in Experiment 4), and 14 ("Apartment 8" in Experiment 4) were used. Apartments 12 and 14 were synthetically generated.

Apartment ID	two-feature model			eight-feature model		
	prediction	error	error fraction	prediction	error	error fraction
1	1,015,000	90,000	0.097	957,560	32,560	0.035
2	1,170,000	-80,000	0.064	1,166,040	-83,960	0.067
3	1,970,000	-560,000	0.221	1,989,200	-540,800	0.214
4	1,580,000	-170,000	0.097	1,573,970	-176,030	0.100
5	630,000	74,000	0.133	634,830	78,830	0.142
6	630,000	-145,000	0.187	555,190	-219,810	0.284
7	1,330,000	331,000	0.331	1,274,700	275,700	0.276
8	1,680,000	-20,000	0.012	1,685,340	-14,660	0.009
9	1,340,000	-210,000	0.135	1,264,600	-285,400	0.184
10	622,000	97,000	0.185	642,820	117,820	0.224
11	1,190,000	541,000	0.834	1,099,550	450,550	0.694
12	1,516,000	—	—	1,475,960	—	—
13	878,000	-292,000	0.25	858,270	-311,730	0.266
14	1,140,000	—	—	1,115,300	—	—

Table 4: Prediction, prediction error (i.e., $m - a$), and prediction error fraction (i.e., $(m - a)/a$) of the two- and eight-feature models on the apartments used in the testing phase of our experiments.

.3 Appendix: Full Distributions of Participant Predictions

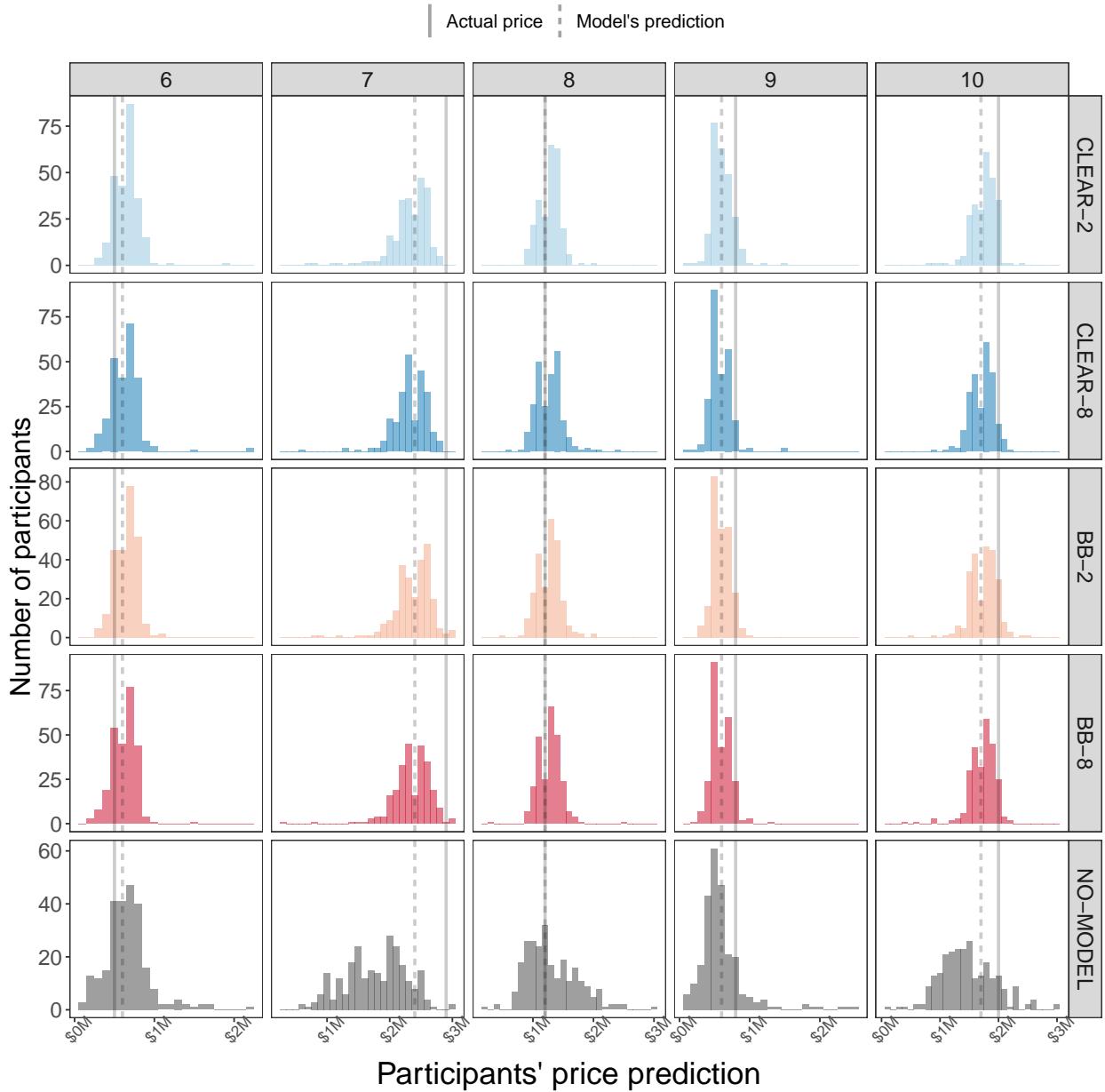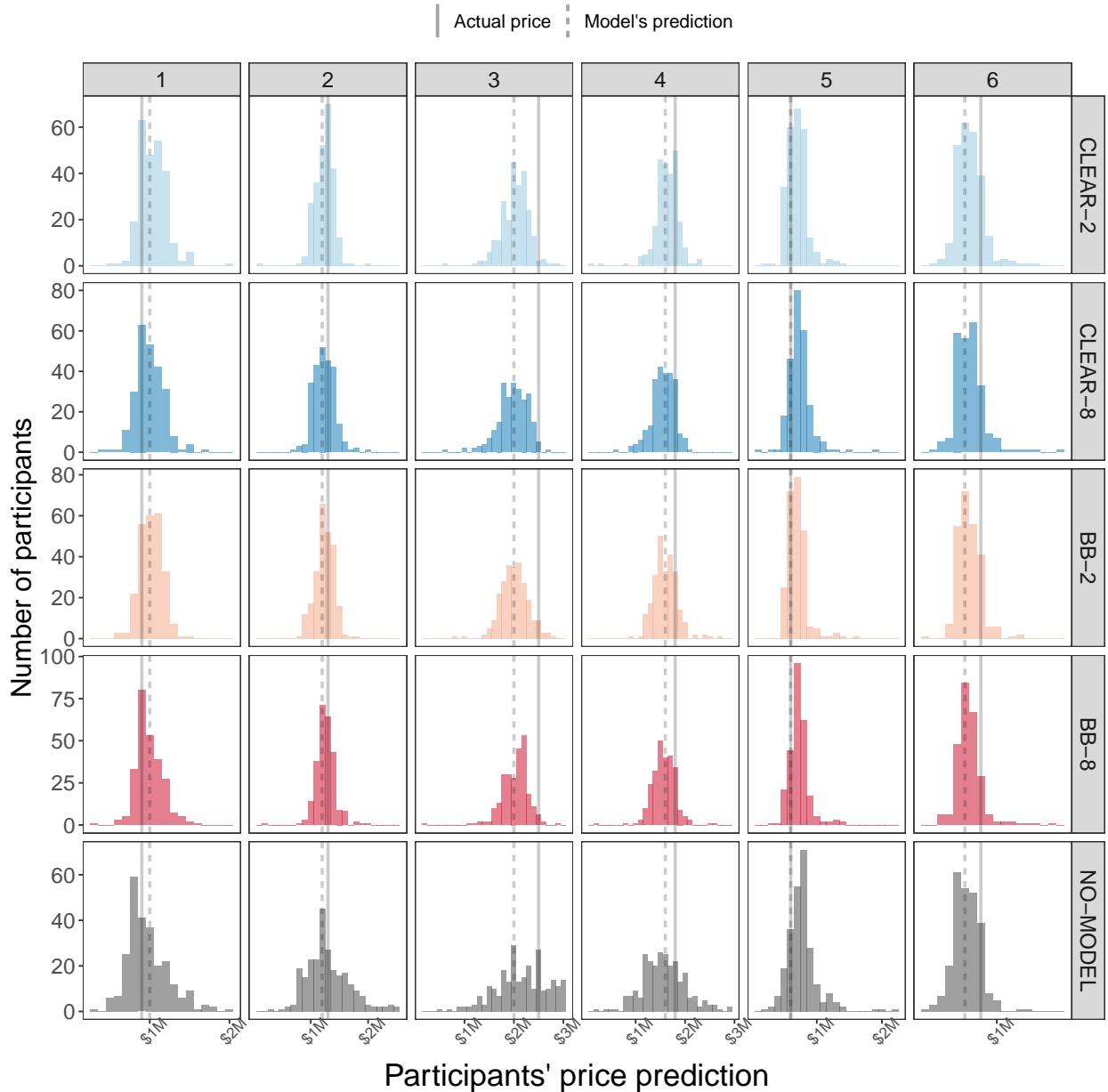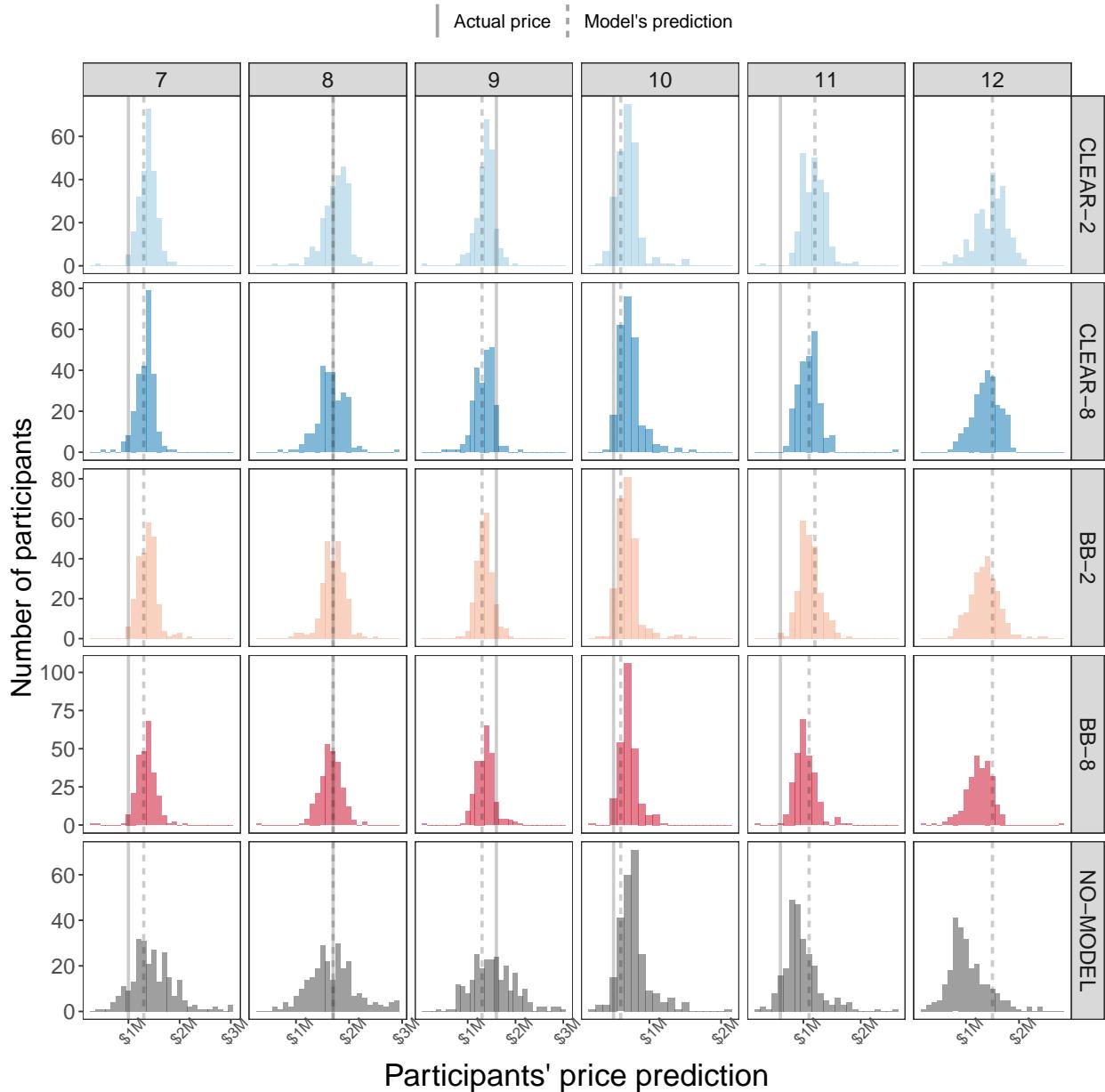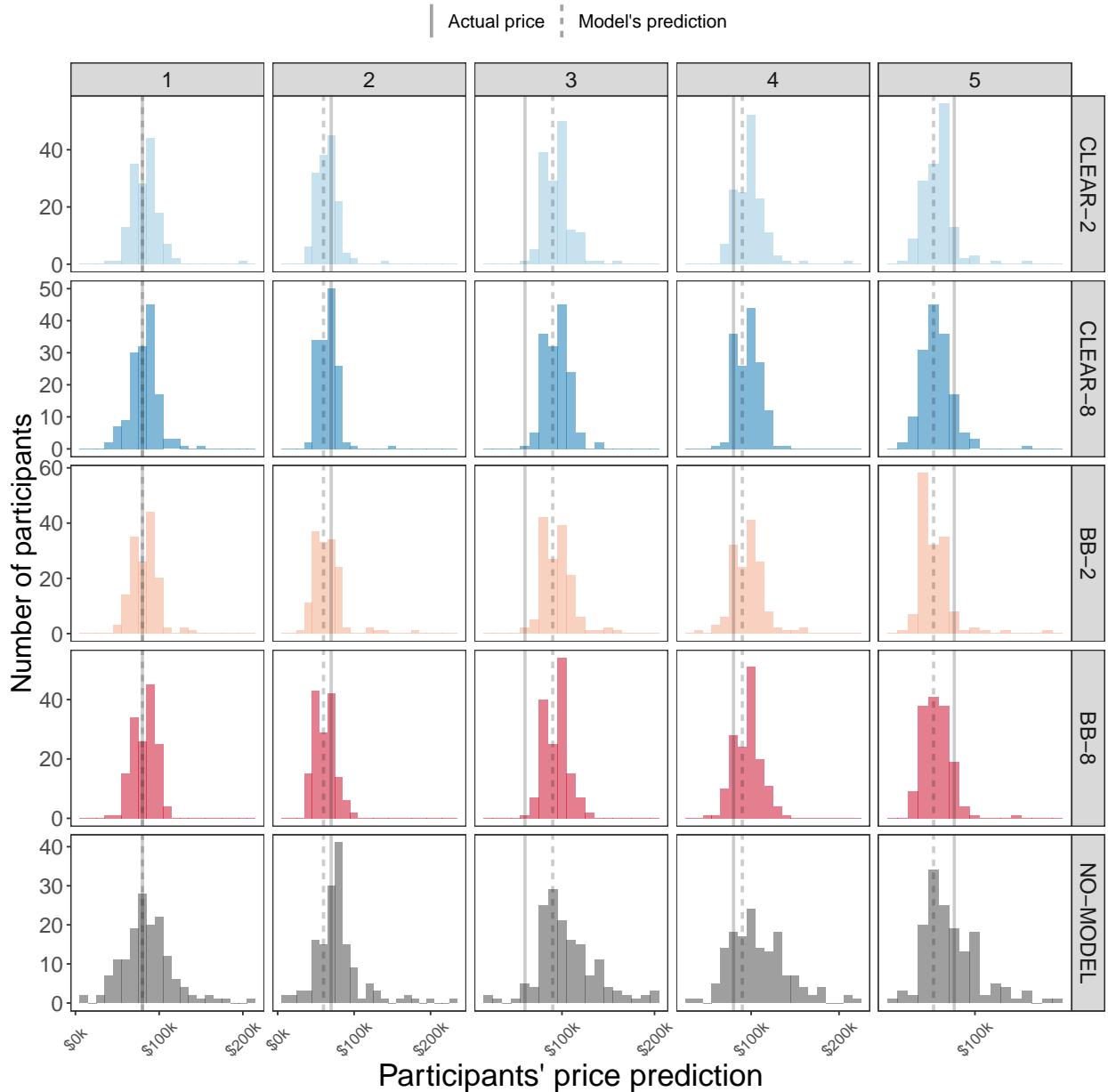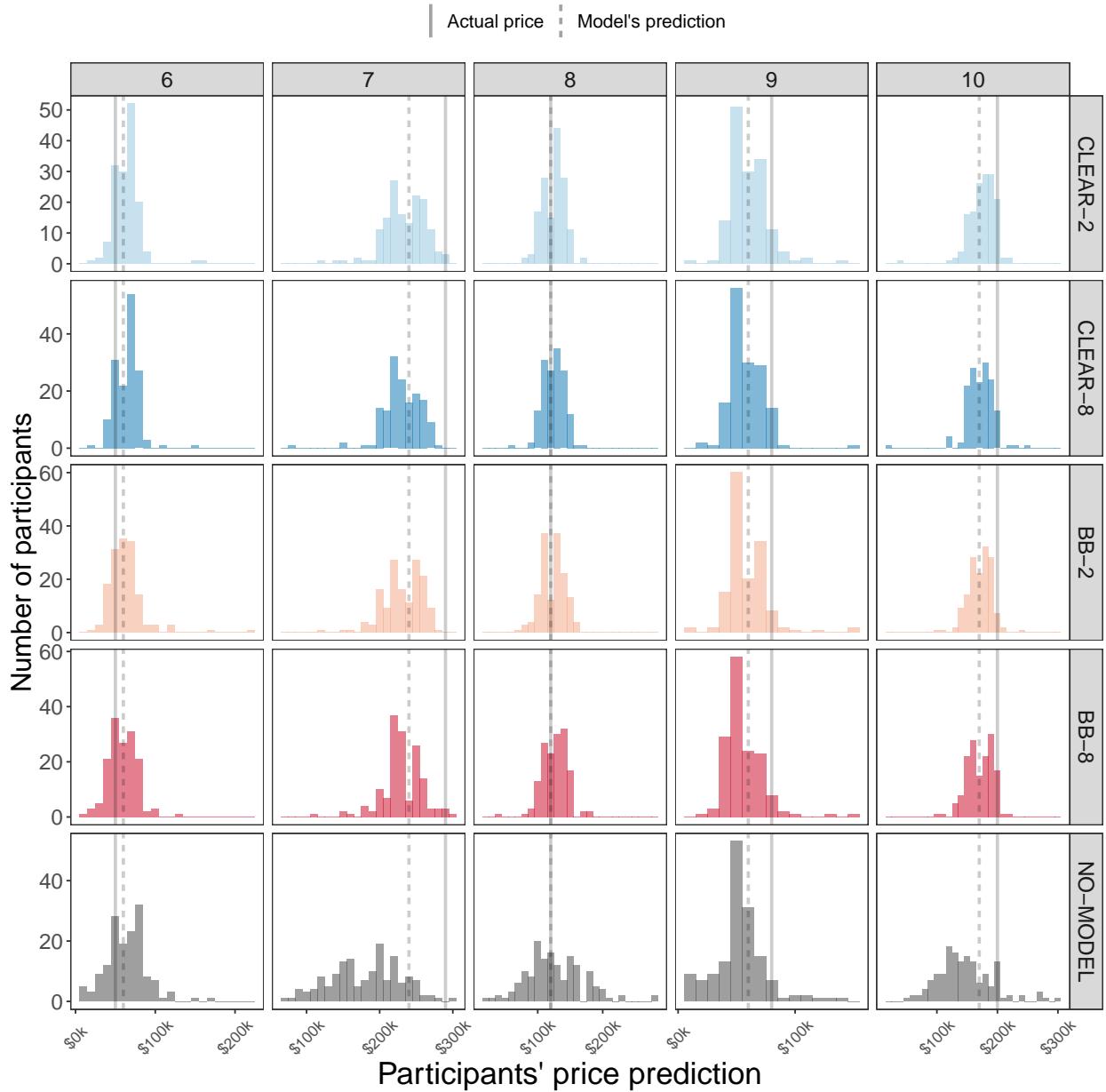Figure 11: Distribution of participants' predictions of prices of apartments 1–5 in the training phase in Experiment 1.

Figure 12: Distribution of participants' predictions of prices of apartments 6–10 in the training phase in Experiment 1.

Figure 13: Distribution of participants' predictions of prices of apartments 1–6 in the testing phase in Experiment 1.

Figure 14: Distribution of participants' predictions of prices of apartments 7–12 in the testing phase in Experiment 1.

Figure 15: Distribution of participants' predictions of prices of apartments 1–5 in the training phase in Experiment 2.

Figure 16: Distribution of participants' predictions of prices of apartments 6–10 in the training phase in Experiment 2.
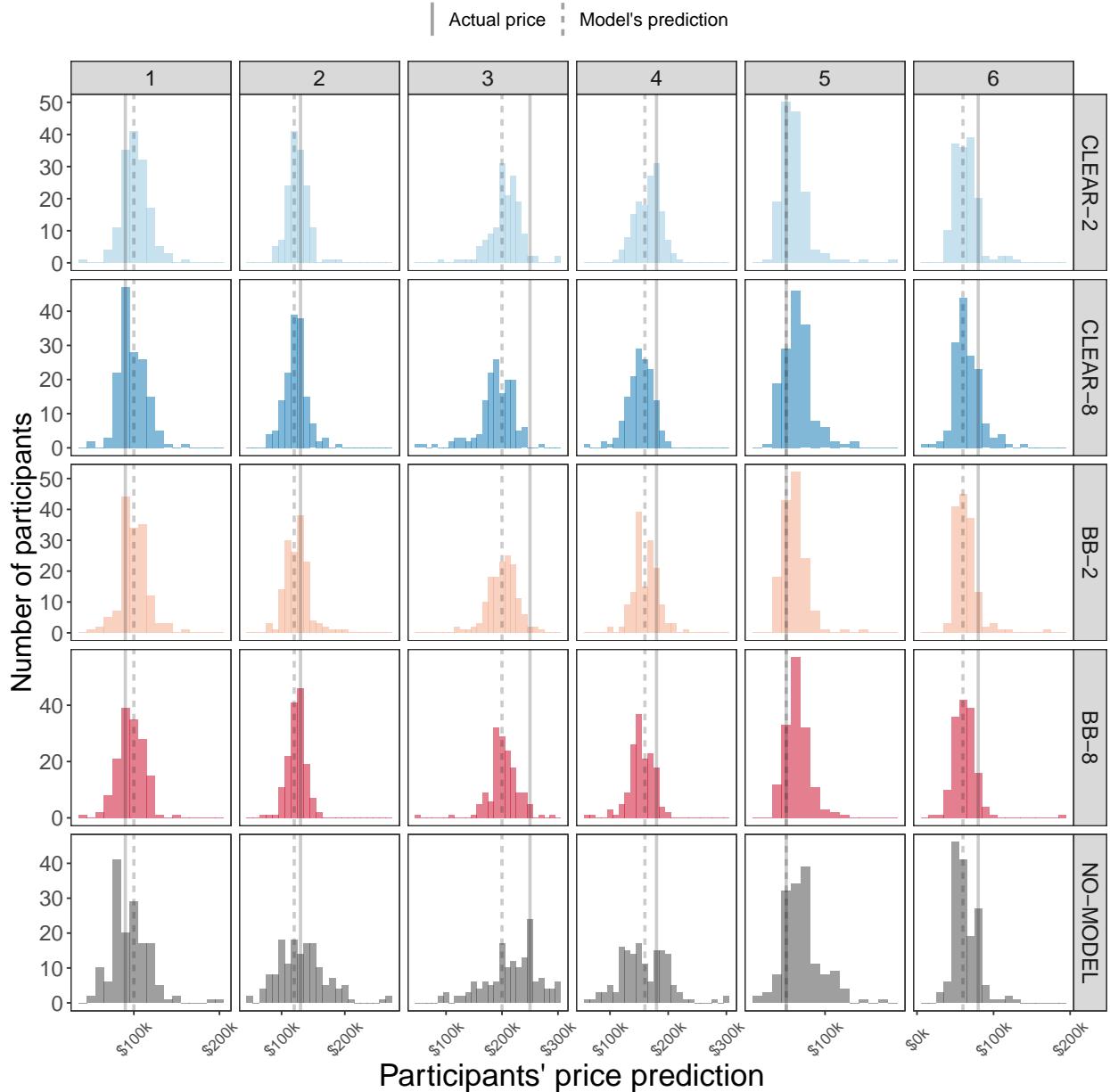
Figure 17: Distribution of participants' predictions of prices of apartments 1–6 in the testing phase in Experiment 2.

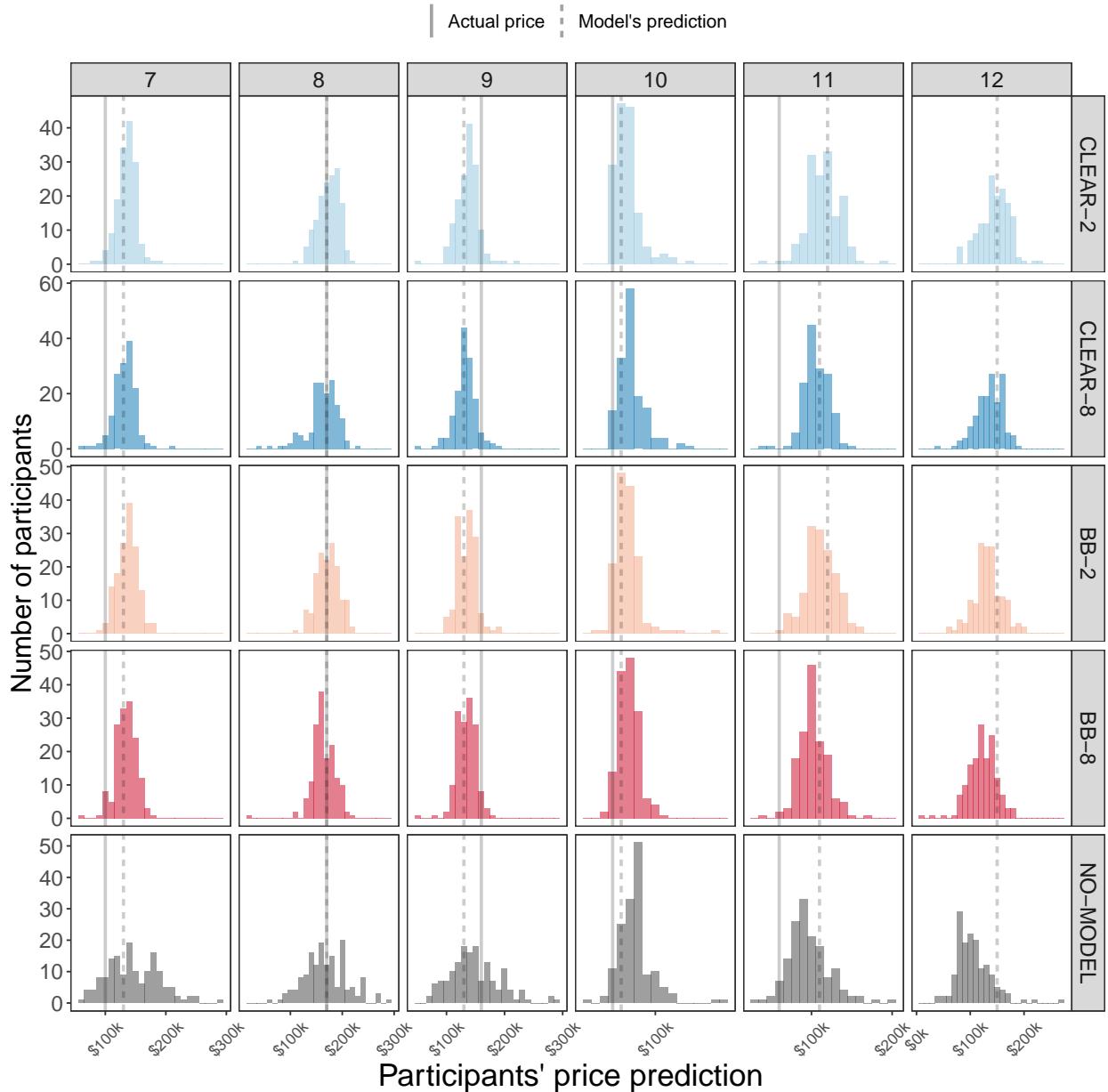# Experiment 2: Representative U.S. prices (testing phase)

Figure 18: Distribution of participants' predictions of prices of apartments 7–12 in the testing phase in Experiment 2.

Figure 19: Distribution of participants' initial (before seeing the model's prediction, dotted) predictions and their final (after seeing the model's prediction, solid) prediction of prices of apartments 1–6 in Experiment 3. Points show the mean initial and final predictions and the arrow indicates the shift in the mean predictions.
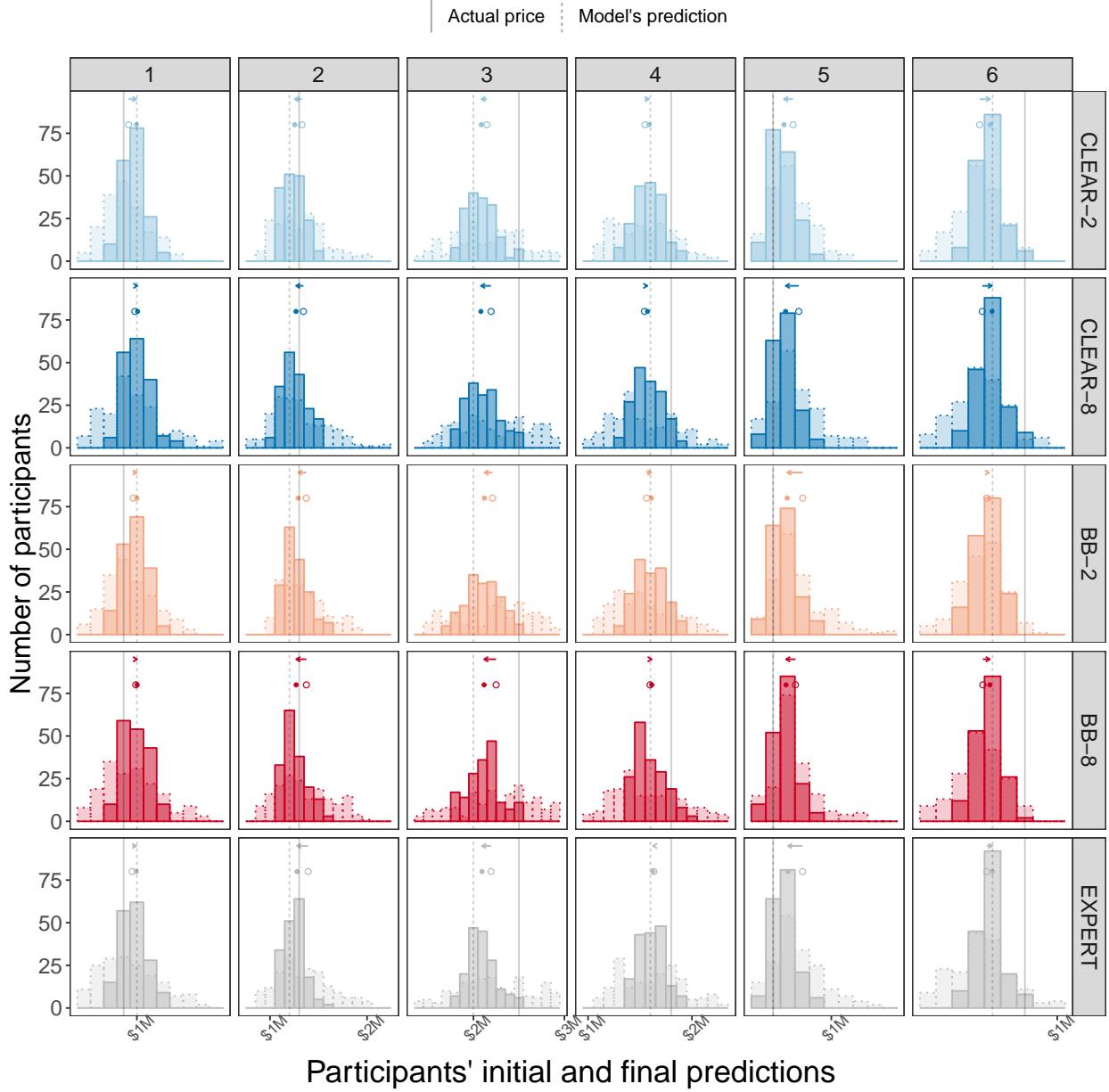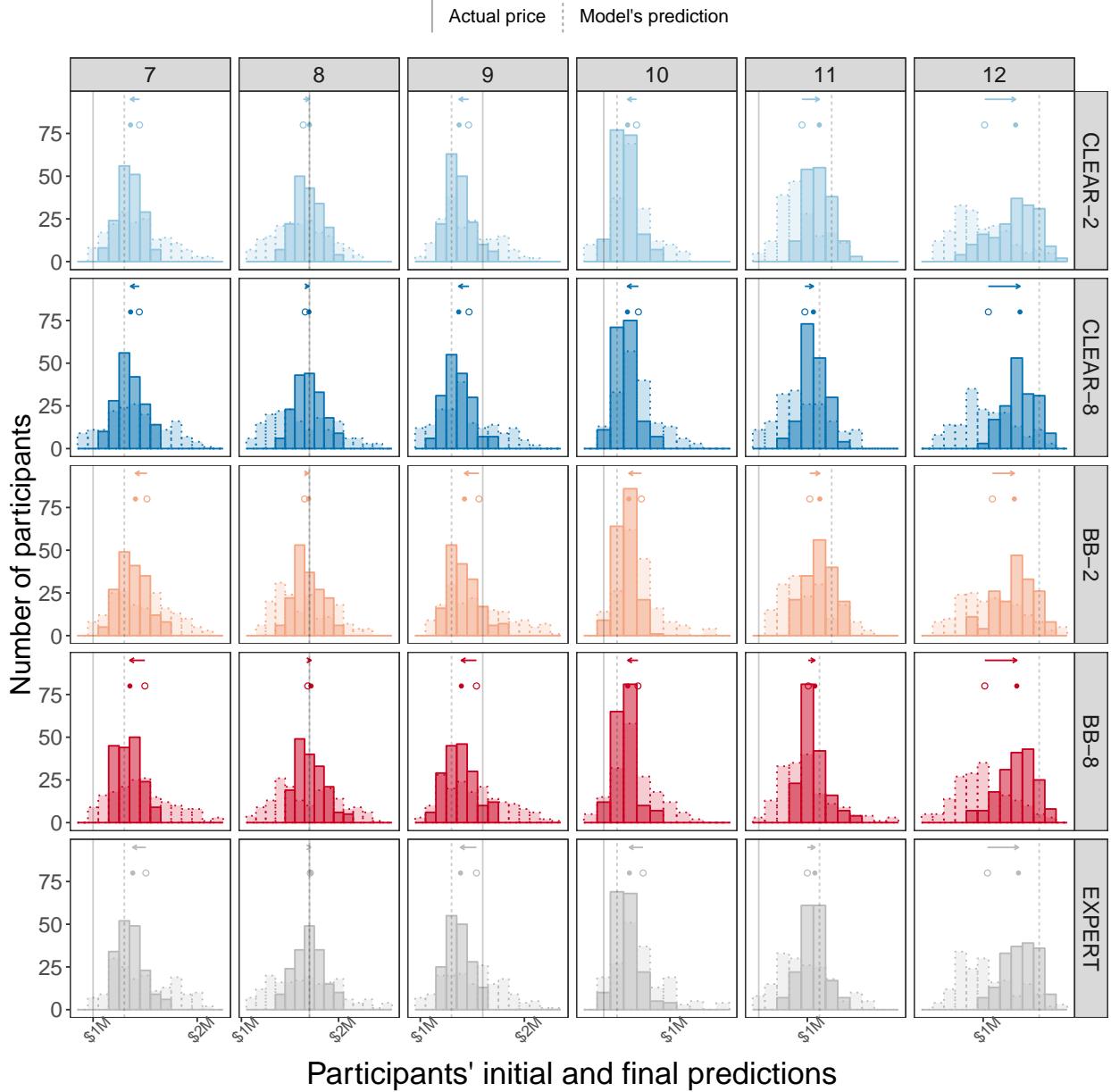
Figure 20: Distribution of participants' initial (before seeing the model's prediction, dotted) predictions and their final (after seeing the model's prediction, solid) prediction of prices of apartments 7–12 in Experiment 3. Points show the mean initial and final predictions and the arrow indicates the shift in the mean predictions.

Experiment 4: Outlier focus (training phase)

Actual price ┆ Model's prediction
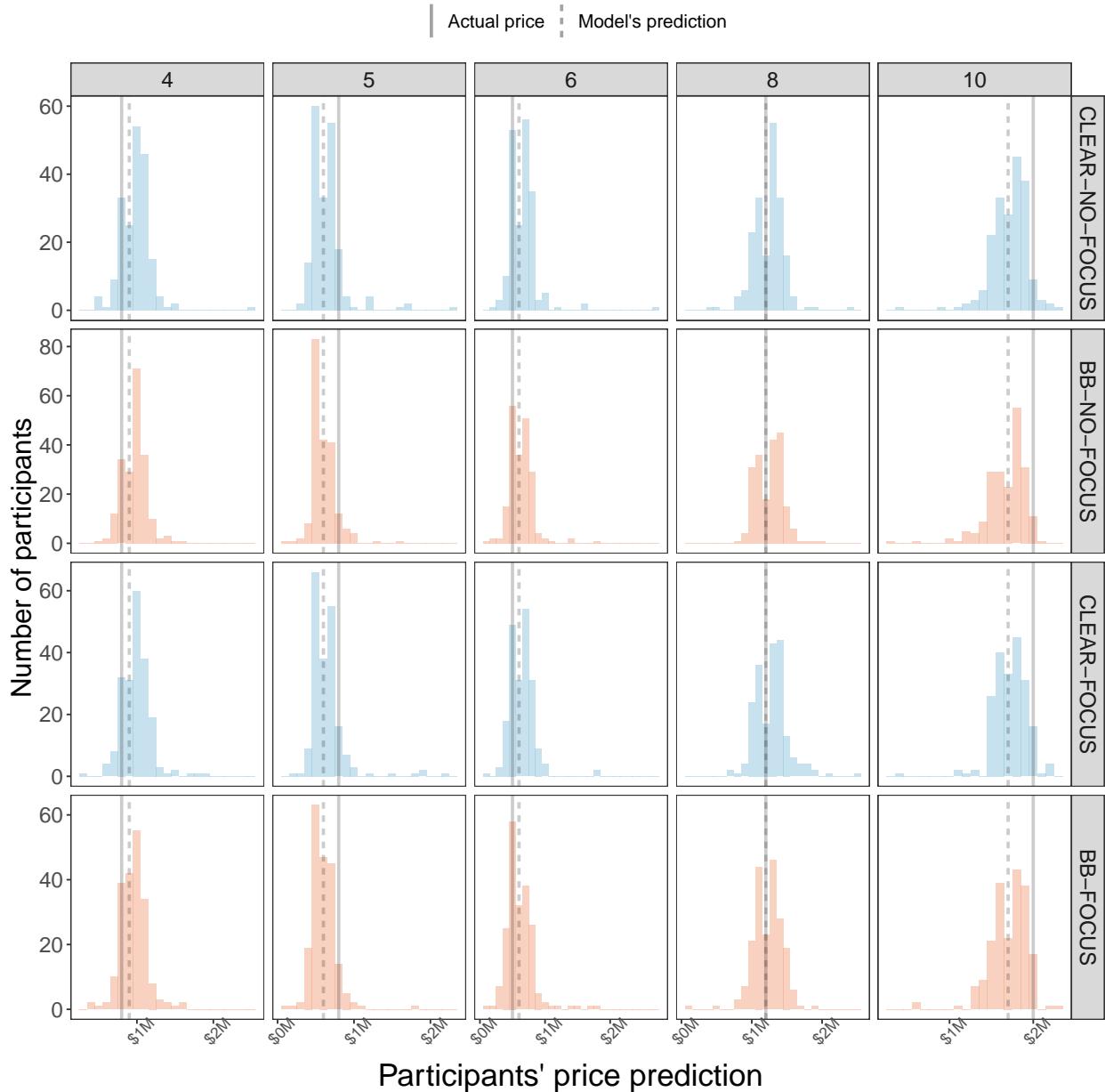

Figure 21: Distribution of participants' predictions of prices of apartments in the training phase in Experiment 4.
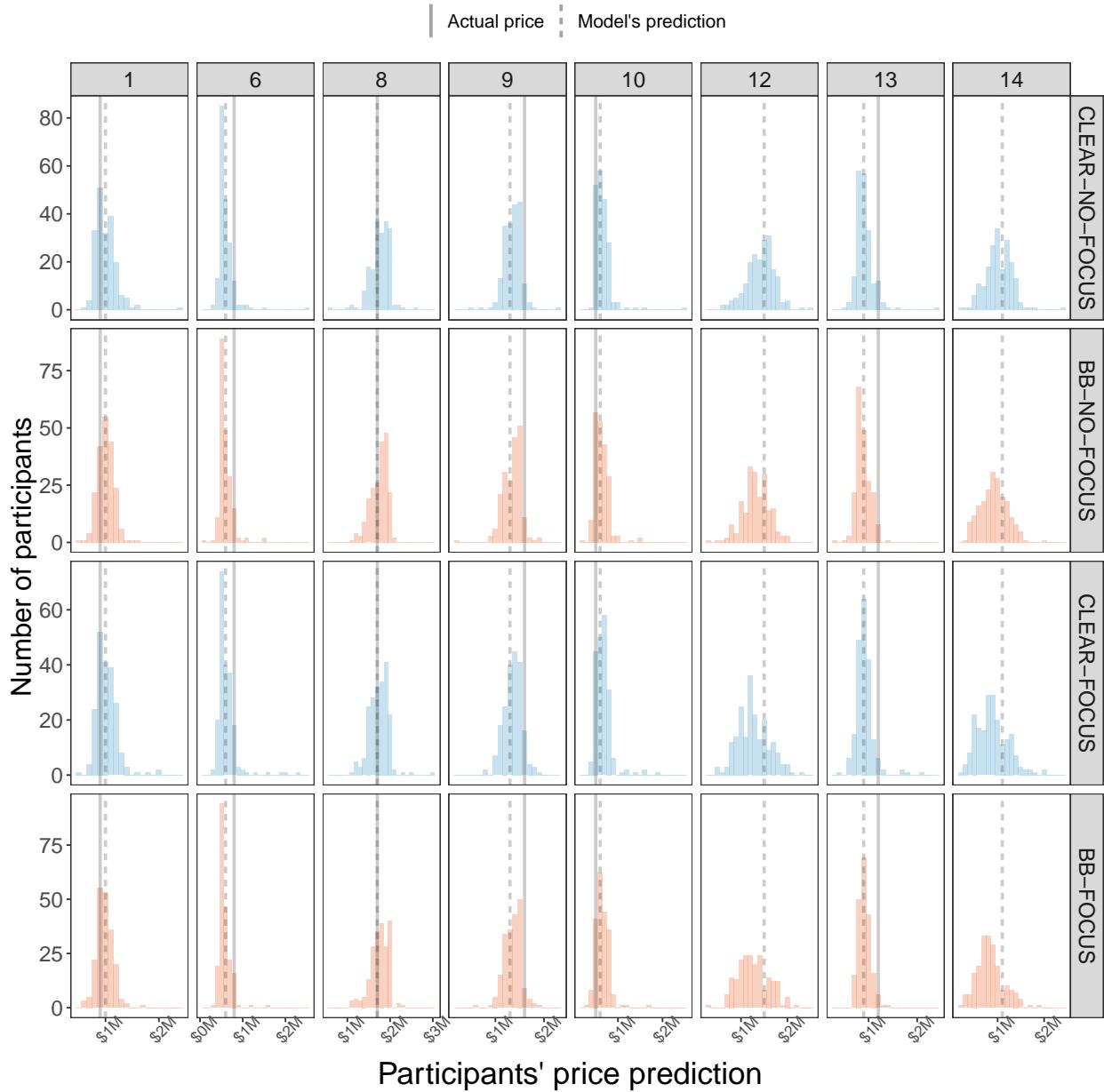
Figure 22: Distribution of participants' predictions of prices of apartments in the testing phase in Experiment 4.

.4 Appendix: ANOVA Tables for Our Experiment

Experiment 1: Predicting Prices

	Sum Sq	Mean Sq	NumDF	DenDF	F value	Pr(>F)
transparency	0.52	0.52	1.00	994.00	12.57	0.0004
num_features	4.90	4.90	1.00	994.00	119.54	0.0000
transparency:num_features	1.70	1.70	1.00	994.00	41.48	0.0000

Table 5: Results from two-way ANOVA on the simulation error in Experiment 1.

	Sum Sq	Mean Sq	NumDF	DenDF	F value	Pr(>F)
transparency	0.10	0.10	1.00	994.00	5.83	0.0159
num_features	0.04	0.04	1.00	994.00	2.15	0.1427
transparency:num_features	0.00	0.00	1.00	994.00	0.06	0.8143

Table 6: Results from two-way ANOVA on the deviation between the model's prediction and participants' prediction of the price in Experiment 1.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
transparency	1	0.02	0.02	1.38	0.2405
num_features	1	0.03	0.03	1.70	0.1920
transparency:num_features	1	0.00	0.00	0.00	0.9509
Residuals	994	17.01	0.02		

Table 7: Results from two-way ANOVA on the deviation between the model's prediction and participants' prediction of the price for apartment 11 in Experiment 1.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
transparency	1	0.34	0.34	8.81	0.0031
num_features	1	0.04	0.04	1.13	0.2882
transparency:num_features	1	0.13	0.13	3.32	0.0687
Residuals	994	38.07	0.04		

Table 8: Results from two-way ANOVA on the deviation between the model's prediction and participants' prediction of the price for apartment 12 in Experiment 1.

Experiment 2: Scaled-down prices

	Sum Sq	Mean Sq	NumDF	DenDF	F value	Pr(>F)
transparency	0.003	0.003	1.00	594.00	7.54	0.0062
num_features	0.032	0.032	1.00	594.00	75.45	0.0000
transparency:num_features	0.018	0.018	1.00	594.00	43.14	0.0000

Table 9: Results from two-way ANOVA on the simulation error in Experiment 2.

	Sum Sq	Mean Sq	NumDF	DenDF	F value	Pr(>F)
transparency	0.001	0.001	1.00	594.00	3.33	0.0685
num_features	0.000	0.000	1.00	594.00	1.29	0.2556
transparency:num_features	0.000	0.000	1.00	594.00	1.82	0.1775

Table 10: Results from two-way ANOVA on the deviation between the model's prediction and participants' prediction of the price in Experiment 2.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
transparency	1.0	0.001	0.001	3.29	0.0702
num_features	1.0	0.001	0.001	4.51	0.0340
transparency:num_features	1.0	0.000	0.000	1.20	0.2731
Residuals	594.0	0.092	0.000		

Table 11: Results from two-way ANOVA on the deviation between the model's prediction and participants' prediction of the price for apartment 11 in Experiment 2.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
transparency	1.0	0.007	0.007	17.53	0.0000
num_features	1.0	0.002	0.002	4.05	0.0446
transparency:num_features	1.0	0.001	0.001	2.31	0.1291
Residuals	594.0	0.229	0.000		

Table 12: Results from two-way ANOVA on the deviation between the model's prediction and participants' prediction of the price for apartment 12 in Experiment 2.

Experiment 3: Weight of Advice

	Sum Sq	Mean Sq	NumDF	DenDF	F value	Pr(>F)
transparency	0.07	0.07	1.00	798.00	4.20	0.0409
num_features	0.01	0.01	1.00	798.00	0.66	0.4151
transparency:num_features	0.02	0.02	1.00	798.00	1.20	0.2731

Table 13: Results from two-way ANOVA on the deviation between the model's prediction and participants' prediction of the price in the four primary conditions in Experiment 3.

	Sum Sq	Mean Sq	NumDF	DenDF	F value	Pr(>F)
condition	0.09	0.02	4.00	994.00	1.45	0.2147

Table 14: Results from one-way ANOVA on the deviation between the model's prediction and participants' prediction of the price in all conditions (including the "human expert" condition) in Experiment 3.

	Sum Sq	Mean Sq	NumDF	DenDF	F value	Pr(>F)
transparency	2.14	2.14	1.00	817.77	10.47	0.0013
num_features	0.42	0.42	1.00	817.77	2.07	0.1509
transparency:num_features	0.32	0.32	1.00	817.77	1.57	0.2109

Table 15: Results from two-way ANOVA on the weight of advice in the four primary conditions in Experiment 3.

	Sum Sq	Mean Sq	NumDF	DenDF	F value	Pr(>F)
condition	2.92	0.73	4.00	1013.65	3.77	0.0048

Table 16: Results from one-way ANOVA on weight of advice in all conditions (including the "human expert" condition) in Experiment 3.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
transparency	1	0.01	0.01	0.92	0.3380
num_features	1	0.15	0.15	10.87	0.0010
transparency:num_features	1	0.03	0.03	2.08	0.1497
Residuals	798	10.86	0.01		

Table 17: Results from two-way ANOVA on the deviation between the model's prediction and participants' prediction of the price for apartment 11 in Experiment 3.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
transparency	1	0.00	0.00	0.04	0.8439
num_features	1	0.38	0.38	10.26	0.0014
transparency:num_features	1	0.01	0.01	0.35	0.5516
Residuals	798	29.32	0.04		

Table 18: Results from two-way ANOVA on the deviation between the model's prediction and participants' prediction of the price for apartment 12 in Experiment 3.