Exploring the Effects of Machine Learning Literacy Interventions on Laypeople's Reliance on Machine Learning Models

Supplementary Materials

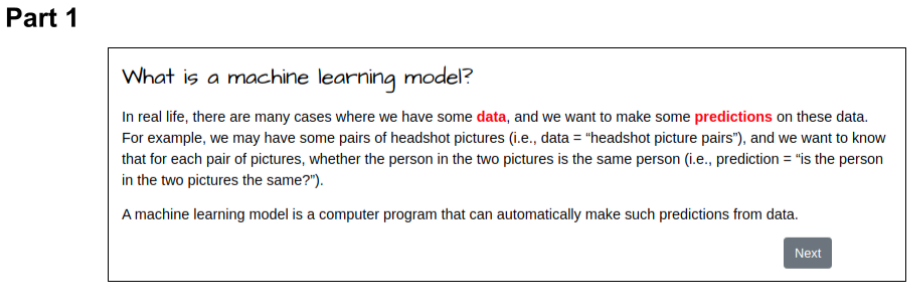
CHUN-WEI CHIANG, Purdue University, USA
MING YIN, Purdue University, USA

ACM Reference Format:
Chun-Wei Chiang and Ming Yin. 2022. Exploring the Effects of Machine Learning Literacy Interventions on Laypeople's Reliance on Machine Learning Models: Supplementary Materials. In *27th International Conference on Intelligent User Interfaces (IUI '22), March 22–25, 2022, Helsinki, Finland.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3490099.3511121

In Sections 1–4, we provide the interface of the user tutorial that we designed to increase subjects' ML literacy. All user tutorials of our experiment always contain the same five parts. Part 1 introduces the purpose of ML models; Part 2 addressees the training process of ML models; Part 3 discusses the evaluation process of the ML models; Part 4 emphasizes the potential systematic performance disparities of ML models when evaluated on different data; and Part 5 explains that the systematic performance disparity could be partly caused by the composition differences across the model's training and evaluation datasets. In Sections 5–7, we provide the task interfaces that subjects saw during different stages of the experiment.

1 TUTORIAL INTERFACE OF THE GENERAL STATIC TREATMENT

The following user tutorial was shown to the subjects who were assigned to the General Static treatment in our experiment.



Manuscript submitted to ACM

Part 2

How to get a machine learning model?

We build a machine learning model using **training data**, which is a set of data for which the outcome we are interested in predicting is ***already known***. For example, to build a machine learning model to predict whether the person in two pictures is the same person or not, the training data could be a set of headshot picture pairs, and each pair is associated with a ***label*** indicating whether the person in the two pictures is the same person or not.

"Training" a machine learning model, then, is having the computer "look at" the training data and identify the **patterns** that will inform the predictions---for example, in a pair of headshot pictures, if one person has a square face while the other has a round face, they are unlikely to be the same person. The machine learning model effectively summarizes all the patterns that the computer has detected from the training data.

Next

Part 3

How good is a machine learning model?

We can evaluate the quality of a machine learning model by using the model to make predictions on some test datasets, and verify whether those predictions are correct.

A widely-used metric of model performance is **accuracy**, that is, the proportion of correct predictions a machine learning model makes during the evaluation. Intuitively, the higher the accuracy, the better the model is.

Next

Part 4

Model performance can be different on different data

Machine learning models can exhibit **systematic performance disparities** when evaluated on different data. Again, consider face recognition systems that are often powered by machine learning models to predict whether the persons in two pictures are the same person.

In a paper published in October 2019 [1], researchers found that commercial face recognition systems' error rates on African faces are twice of the error rates on Caucasians, as shown in the following table.

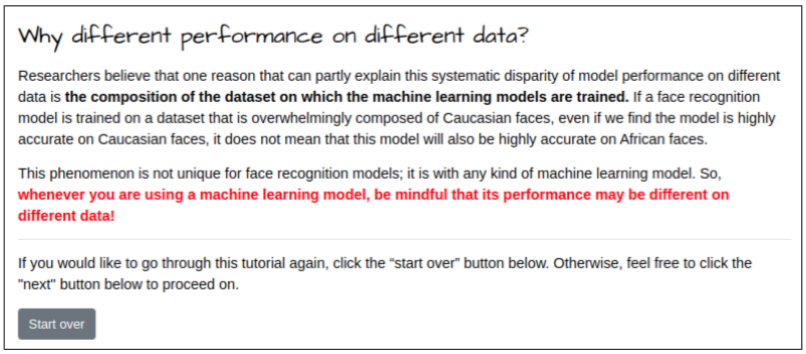
Face detection system / Accuracy(%)	Caucasians 	Indians 	East Asians 	Africans 
Microsoft	87.60	82.83	79.67	75.83
Face++	93.90	88.55	92.47	87.50
Baidu	89.13	86.53	90.27	77.97
Amazon	90.45	87.20	84.87	86.37

(For each system, its best-performing demographic group is shown in green, and its worst-performing demographic group is shown in red.)

[1] Wang, Mei, et al. "Racial faces in the wild: Reducing racial bias by information maximization adaptation network." Proceedings of the IEEE International Conference on Computer Vision. 2019.

Next

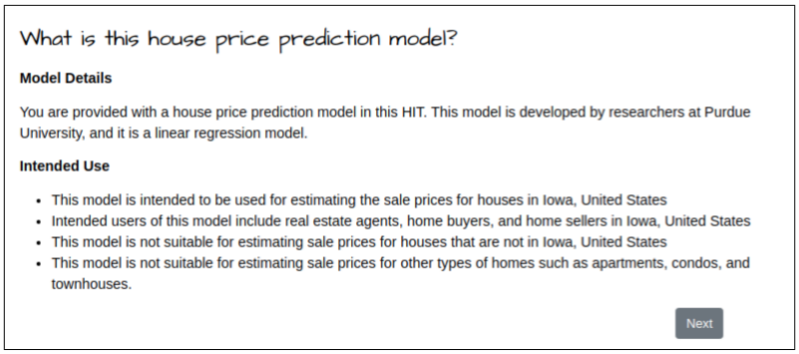
Part 5


Why different performance on different data?
Researchers believe that one reason that can partly explain this systematic disparity of model performance on different data is the composition of the dataset on which the machine learning models are trained. If a face recognition model is trained on a dataset that is overwhelmingly composed of Caucasian faces, even if we find the model is highly accurate on Caucasian faces, it does not mean that this model will also be highly accurate on African faces.
This phenomenon is not unique for face recognition models; it is with any kind of machine learning model. So, whenever you are using a machine learning model, be mindful that its performance may be different on different data!
If you would like to go through this tutorial again, click the "start over" button below. Otherwise, feel free to click the "next" button below to proceed on.
Start over

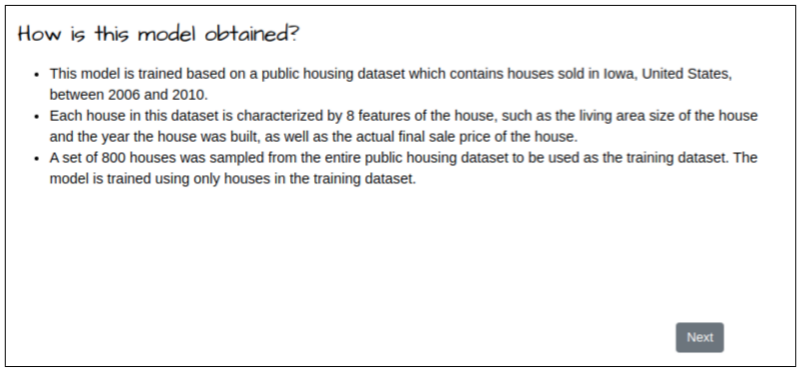
2 TUTORIAL INTERFACE OF THE SPECIFIC STATIC TREATMENT

The following user tutorial was shown to the subjects who were assigned to the Specific Static treatment in our experiment.

Part 1



Part 2



Part 3

How good is this model?

Evaluation Data

A set of 200 houses was sampled from the same public housing dataset to be used as the test dataset (houses that have been selected into the training dataset are excluded from the test dataset). The model's performance is evaluated on this test dataset.

Factors

- Factors that may influence the model performance that have been considered in this evaluation include the living area size of the house and the finish quality of the house.

Metrics

- The model's performance is evaluated using **absolute percentage errors (APE)**, which is the absolute difference between the predicted price and the actual sale price of a house, divided by the actual price.
- Given a set of test houses, the mean value of APE (MAPE) is reported, and 95% confidence intervals are calculated with bootstrap resampling. Intuitively, the smaller the MAPE, the better the model, and the wider the error bars, the less certain we are about the model's performance.

Next

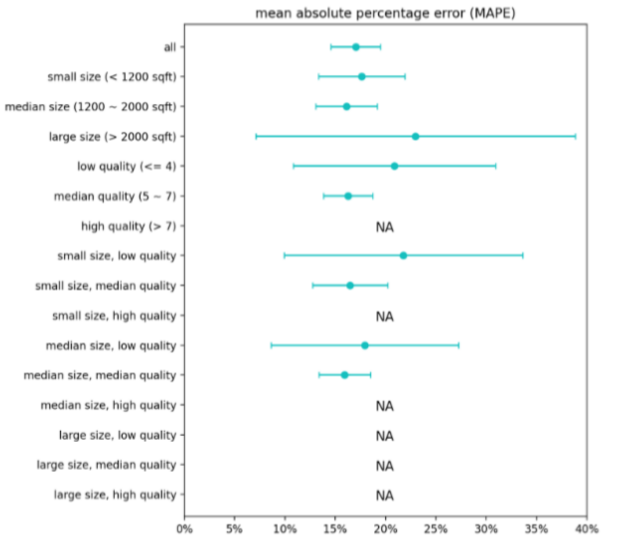
Part 4

How good is this model?

Quantitative Analyses

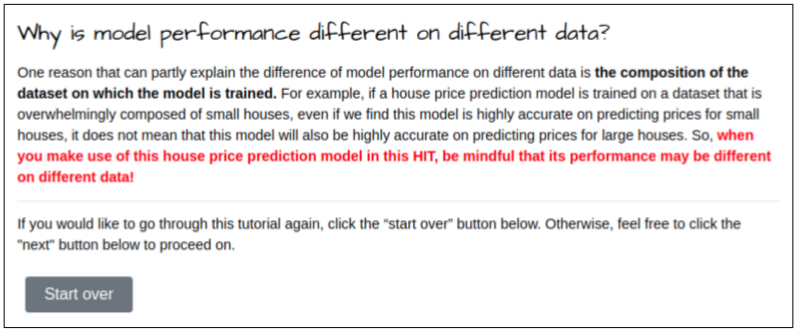
The plot below shows the model's performance on the test dataset. We also break down the test dataset into subsets based on the two factors (size and quality), and show the model's performance on each subset. As you can see, the model's performance is different on different houses.

Note that some subset contains fewer than 5 houses, which is insufficient for us to make a reliable conclusion on the model's performance for houses in that subset. In this case, we indicate the model's performance using "NA".



Next

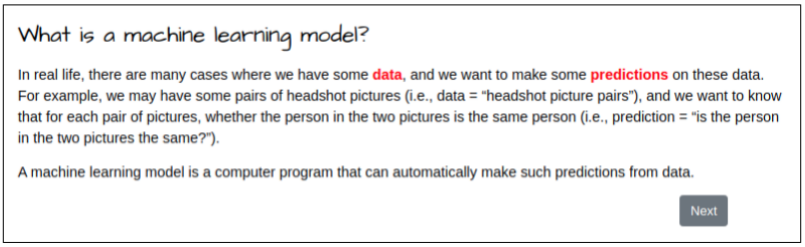
Part 5


Why is model performance different on different data?
One reason that can partly explain the difference of model performance on different data is the composition of the dataset on which the model is trained. For example, if a house price prediction model is trained on a dataset that is overwhelmingly composed of small houses, even if we find this model is highly accurate on predicting prices for small houses, it does not mean that this model will also be highly accurate on predicting prices for large houses. So, when you make use of this house price prediction model in this HIT, be mindful that its performance may be different on different data!
If you would like to go through this tutorial again, click the "start over" button below. Otherwise, feel free to click the "next" button below to proceed on.
Start over

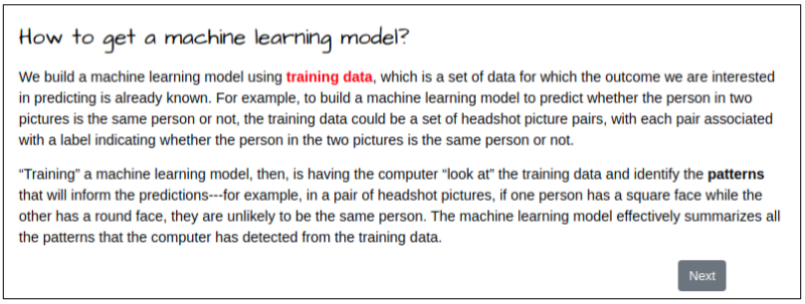
3 TUTORIAL INTERFACE OF THE GENERAL SANDBOX TREATMENT

The following user tutorial was shown to the subjects who were assigned to the General Sandbox treatment in our experiment.

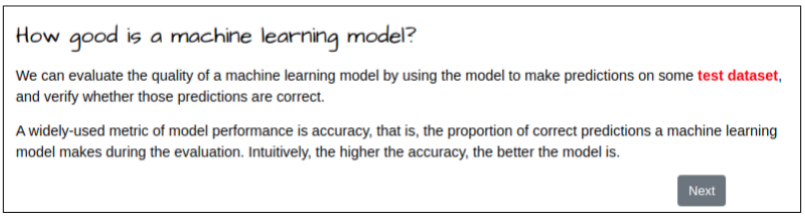
Part 1

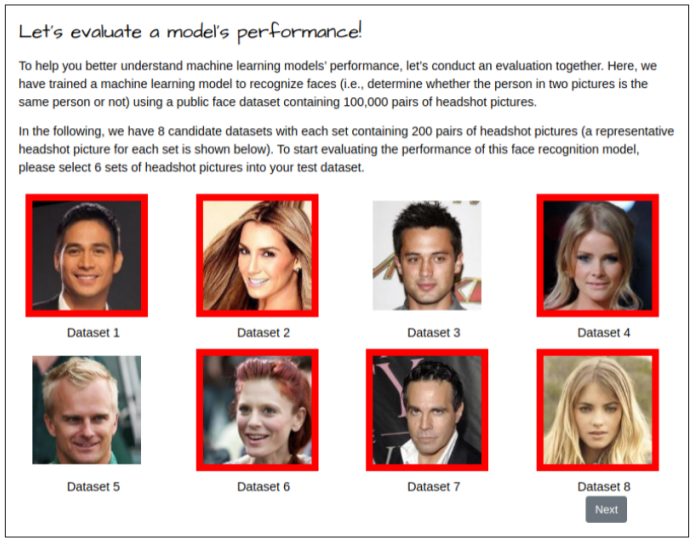


Part 2

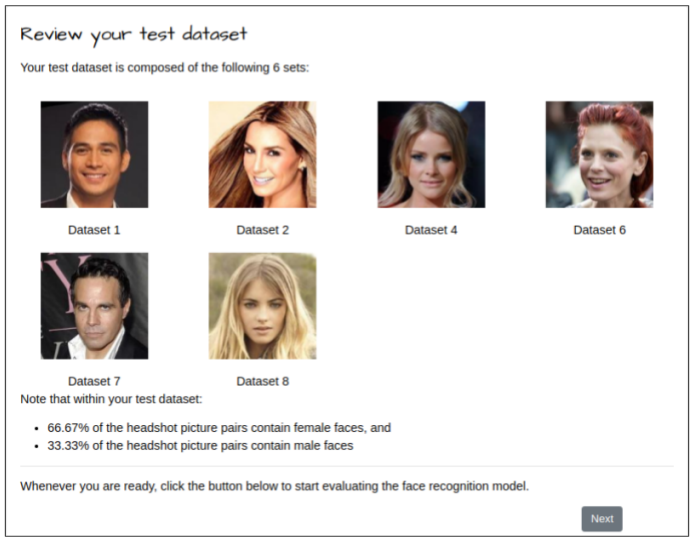


Part 3

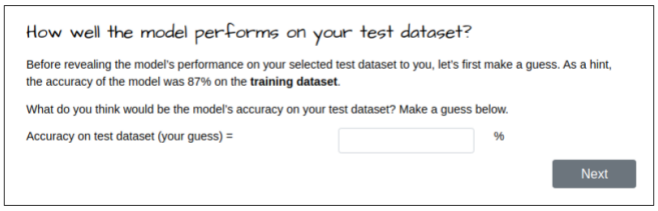




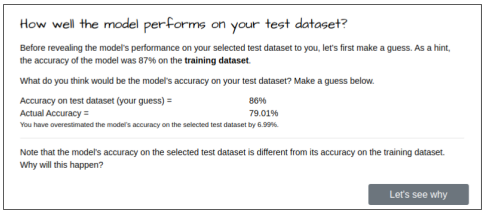
(a) Step 1: Subjects constructed a testing dataset.



(b) Step 2: Subjects were given a summary of the selected testing dataset.



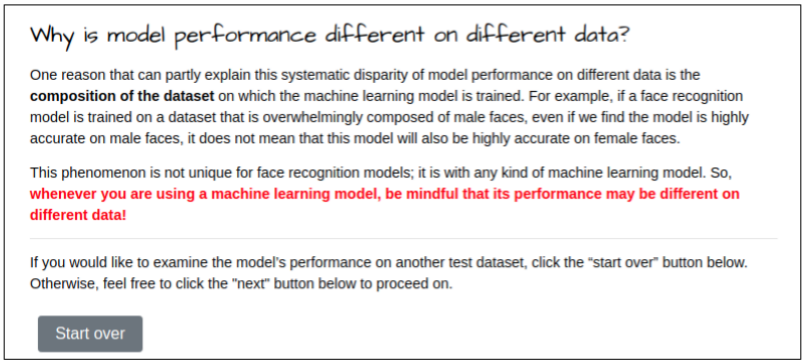
(c) Step 3: Subjects were asked to guess the performance of the model on the selected testing dataset.



(d) Step 4: Subjects were shown the model's performance on the selected testing dataset.

Fig. 1. The interface for Part 4 of the tutorial used in the GENERAL SANDBOX treatment.

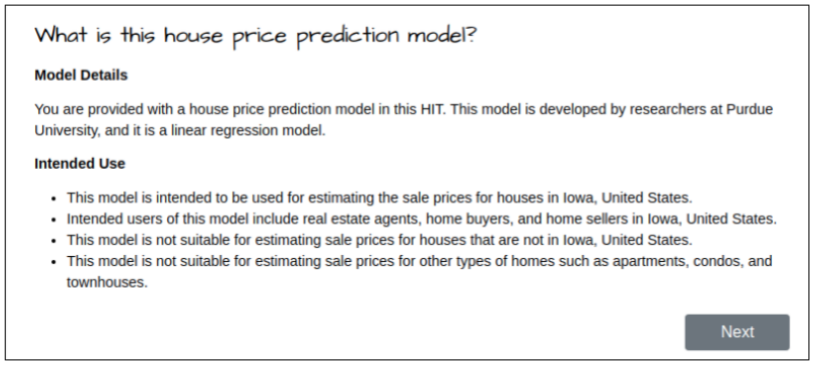
Part 5


Why is model performance different on different data?
One reason that can partly explain this systematic disparity of model performance on different data is the composition of the dataset on which the machine learning model is trained. For example, if a face recognition model is trained on a dataset that is overwhelmingly composed of male faces, even if we find the model is highly accurate on male faces, it does not mean that this model will also be highly accurate on female faces.
This phenomenon is not unique for face recognition models; it is with any kind of machine learning model. So, whenever you are using a machine learning model, be mindful that its performance may be different on different data!
If you would like to examine the model's performance on another test dataset, click the "start over" button below. Otherwise, feel free to click the "next" button below to proceed on.
Start over

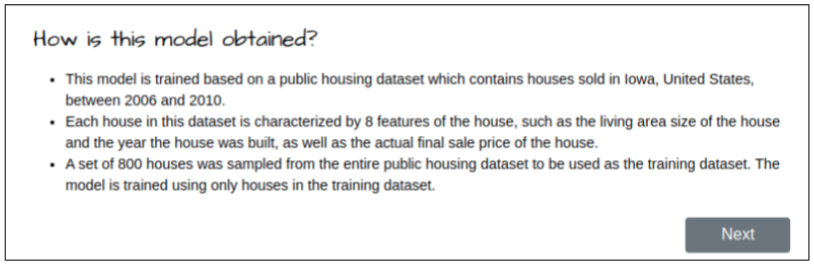
4 TUTORIAL INTERFACE OF THE SPECIFIC SANDBOX TREATMENT

The following user tutorial was shown to the subjects who were assigned to the Specific Sandbox treatment in our experiment.

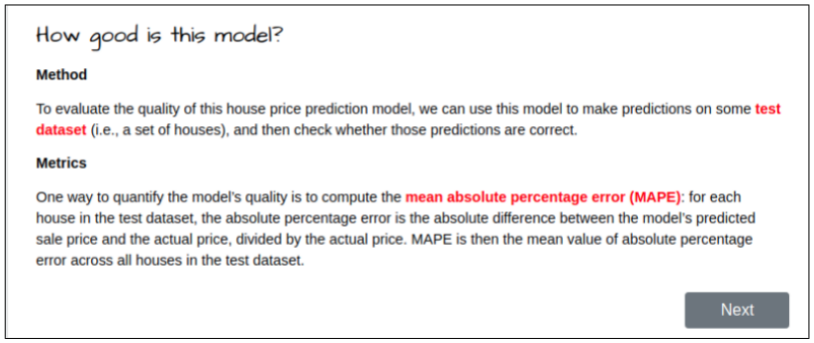
Part 1

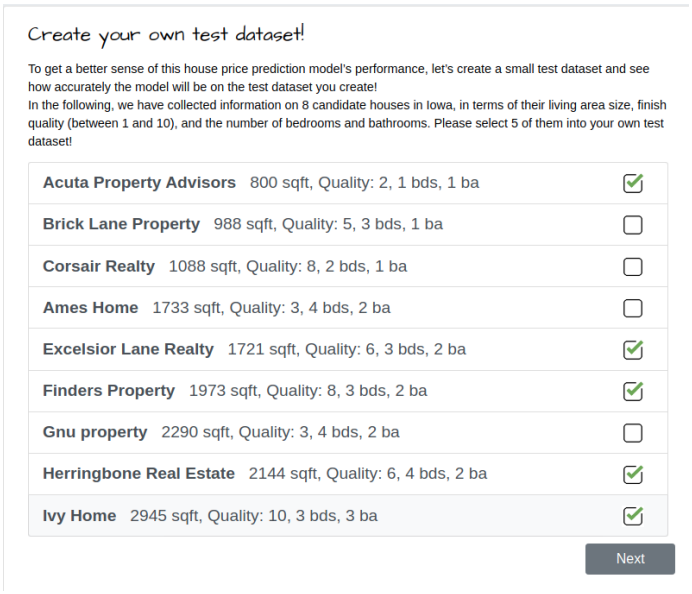


Part 2

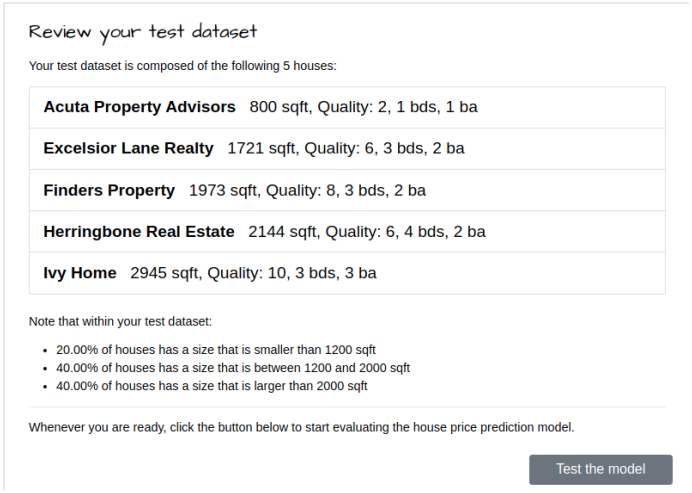


Part 3

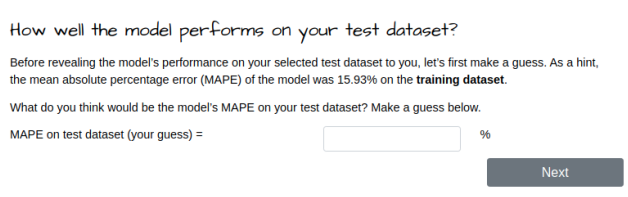




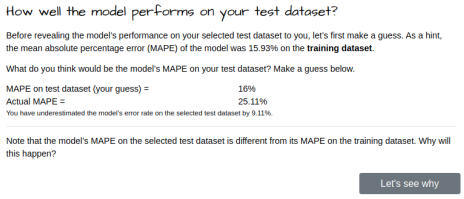
(a) Step 1: Subjects constructed a testing dataset.



(b) Step 2: Subjects were given a summary of the selected testing dataset.



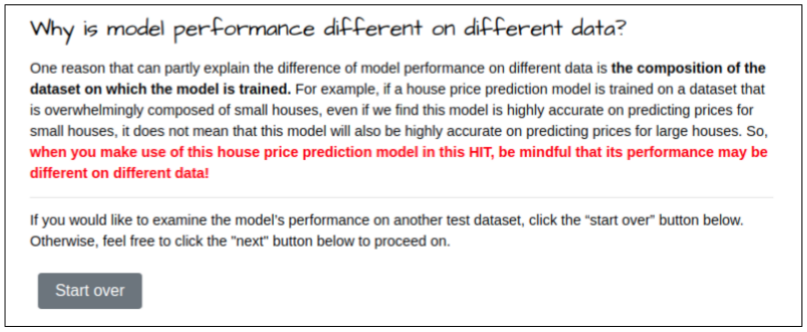
(c) Step 3: Subjects were asked to guess the performance of the model on the selected testing dataset.



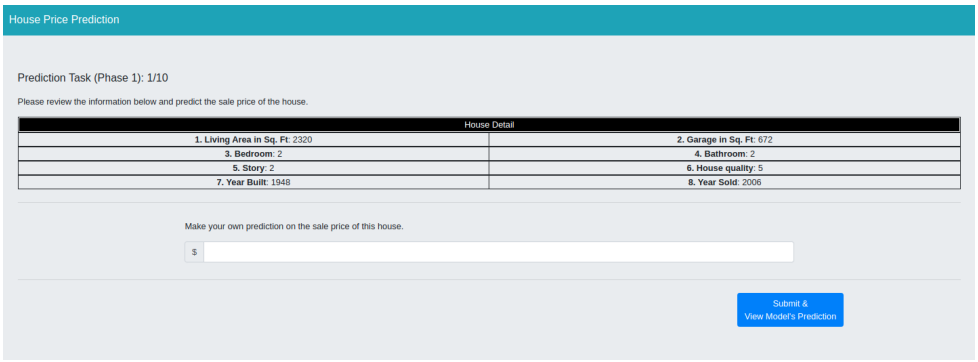
(d) Step 4: Subjects were shown the model's performance on the selected testing dataset.

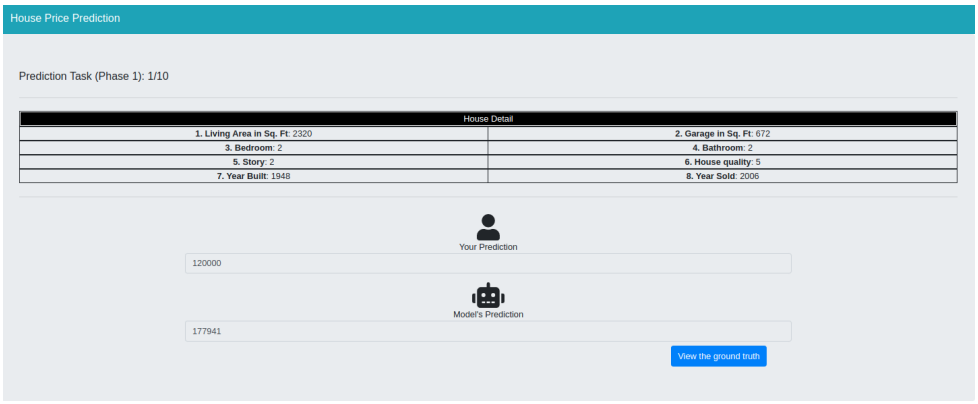
Fig. 2. The interface for Part 4 of the tutorial used in the SPECIFIC SANDBOX treatment.

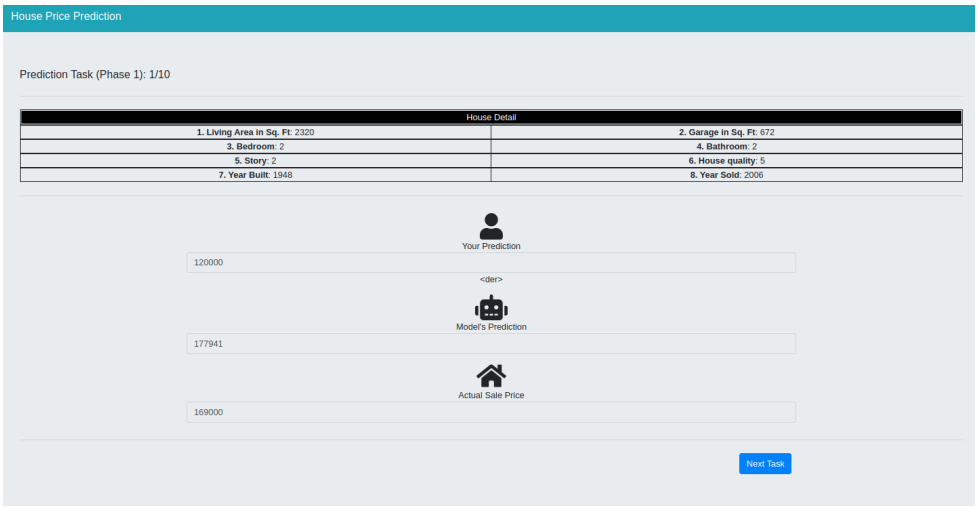
Part 5


Why is model performance different on different data?
One reason that can partly explain the difference of model performance on different data is the composition of the dataset on which the model is trained. For example, if a house price prediction model is trained on a dataset that is overwhelmingly composed of small houses, even if we find this model is highly accurate on predicting prices for small houses, it does not mean that this model will also be highly accurate on predicting prices for large houses. So, when you make use of this house price prediction model in this HIT, be mindful that its performance may be different on different data!
If you would like to examine the model's performance on another test dataset, click the "start over" button below. Otherwise, feel free to click the "next" button below to proceed on.
Start over

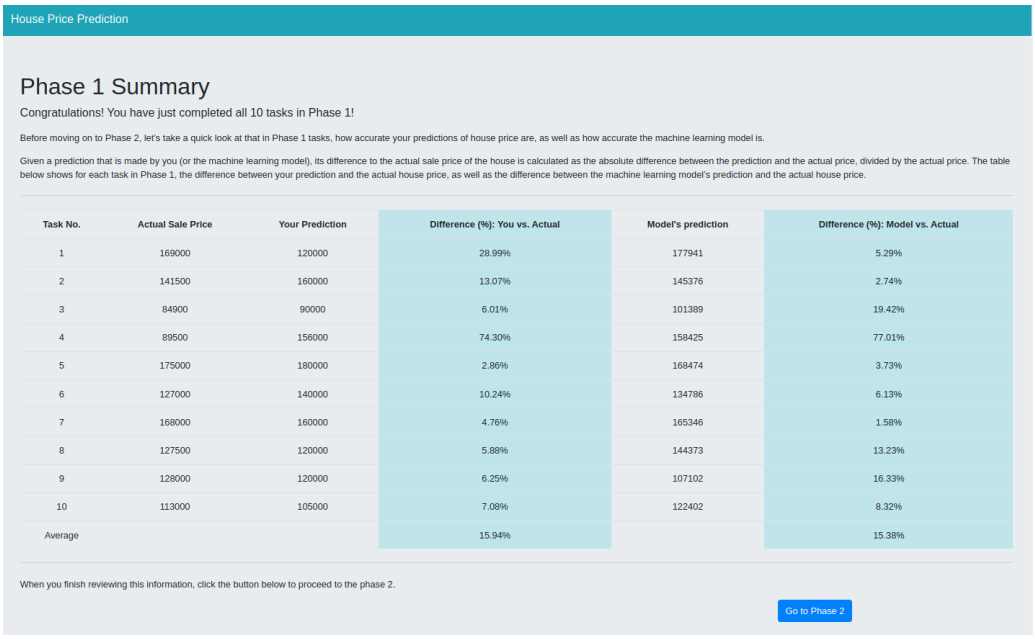
5 THE TASK INTERFACE OF PHASE 1







6 THE INTERFACE OF THE MID-POINT FEEDBACK PAGE



7 THE TASK INTERFACE OF PHASE 2