

# 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.

### Part 1

#### What is a machine learning model?

In real life, there are many cases where we have some **data**, and we want to make some **predictions** on these data. For example, we may have some pairs of headshot pictures (i.e., data = "headshot picture pairs"), and we want to know that for each pair of pictures, whether the person in the two pictures is the same person (i.e., prediction = "is the person in the two pictures the same?").

A machine learning model is a computer program that can automatically make such predictions from data.

[Next](#)

---

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

© 2022 Copyright held by the owner/author(s).

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

What is this house price prediction model?

#### Model Details

You are provided with a house price prediction model in this HIT. This model is developed by researchers at Purdue University, and it is a linear regression model.

#### Intended Use

- This model is intended to be used for estimating the sale prices for houses in Iowa, United States
- Intended users of this model include real estate agents, home buyers, and home sellers in Iowa, United States
- This model is not suitable for estimating sale prices for houses that are not in Iowa, United States
- This model is not suitable for estimating sale prices for other types of homes such as apartments, condos, and townhouses.

Next

### Part 2

How is this model obtained?

- This model is trained based on a public housing dataset which contains houses sold in Iowa, United States, between 2006 and 2010.
- Each house in this dataset is characterized by 8 features of the house, such as the living area size of the house and the year the house was built, as well as the actual final sale price of the house.
- A set of 800 houses was sampled from the entire public housing dataset to be used as the training dataset. The model is trained using only houses in the training dataset.

Next

## 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](#)

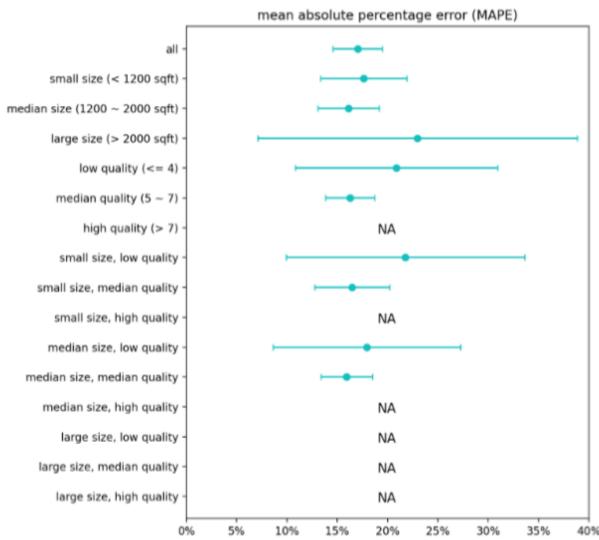
## Part 4

### 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

##### What is a machine learning model?

In real life, there are many cases where we have some **data**, and we want to make some **predictions** on these data. For example, we may have some pairs of headshot pictures (i.e., data = "headshot picture pairs"), and we want to know that for each pair of pictures, whether the person in the two pictures is the same person (i.e., prediction = "is the person in the two pictures the same?").

A machine learning model is a computer program that can automatically make such predictions from data.

Next

#### 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, with each pair 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

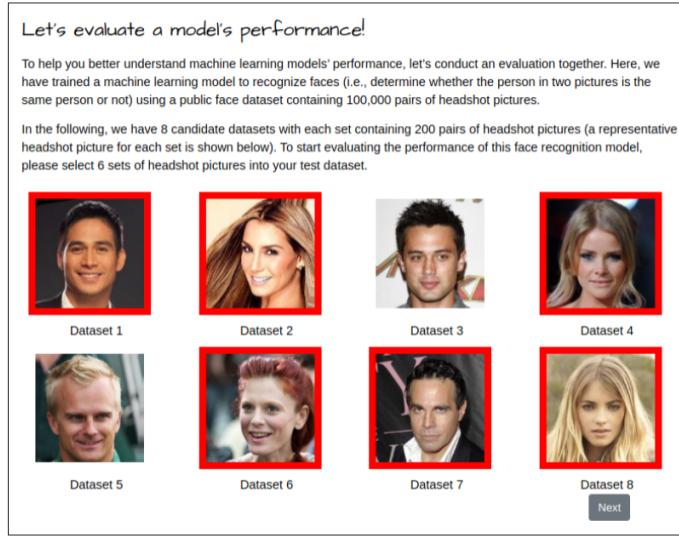
#### Part 3

##### 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 dataset**, 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



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

**Review your test dataset**

Your test dataset is composed of the following 6 sets:

Dataset 1      Dataset 2      Dataset 4      Dataset 6  
 Dataset 7      Dataset 8

Note that within your test dataset:

- 66.67% of the headshot picture pairs contain female faces, and
- 33.33% of the headshot picture pairs contain male faces

Whenever you are ready, click the button below to start evaluating the face recognition model.

**Next**

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

**How well the model performs on your test dataset?**

Before revealing the model's performance on your selected test dataset to you, let's first make a guess. As a hint, the accuracy of the model was 87% on the **training dataset**.

What do you think would be the model's accuracy on your test dataset? Make a guess below.

Accuracy on test dataset (your guess) =  %

**Next**

**How well the model performs on your test dataset?**

Before revealing the model's performance on your selected test dataset to you, let's first make a guess. As a hint, the accuracy of the model was 87% on the **training dataset**.

What do you think would be the model's accuracy on your test dataset? Make a guess below.

Accuracy on test dataset (your guess) =	86%
Actual Accuracy =	79.01%

You have overestimated the model's accuracy on the selected test dataset by 6.99%.

Note that the model's accuracy on the selected test dataset is different from its accuracy on the training dataset. Why will this happen?

**Let's see why**

(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

What is this house price prediction model?

###### Model Details

You are provided with a house price prediction model in this HIT. This model is developed by researchers at Purdue University, and it is a linear regression model.

###### Intended Use

- This model is intended to be used for estimating the sale prices for houses in Iowa, United States.
- Intended users of this model include real estate agents, home buyers, and home sellers in Iowa, United States.
- This model is not suitable for estimating sale prices for houses that are not in Iowa, United States.
- This model is not suitable for estimating sale prices for other types of homes such as apartments, condos, and townhouses.

Next

##### Part 2

How is this model obtained?

- This model is trained based on a public housing dataset which contains houses sold in Iowa, United States, between 2006 and 2010.
- Each house in this dataset is characterized by 8 features of the house, such as the living area size of the house and the year the house was built, as well as the actual final sale price of the house.
- A set of 800 houses was sampled from the entire public housing dataset to be used as the training dataset. The model is trained using only houses in the training dataset.

Next

##### Part 3

How good is this model?

###### Method

To evaluate the quality of this house price prediction model, we can use this model to make predictions on some **test dataset** (i.e., a set of houses), and then check whether those predictions are correct.

###### Metrics

One way to quantify the model's quality is to compute the **mean absolute percentage error (MAPE)**: for each house in the test dataset, the absolute percentage error is the absolute difference between the model's predicted sale price and the actual price, divided by the actual price. MAPE is then the mean value of absolute percentage error across all houses in the test dataset.

Next

Create your own test dataset!

To get a better sense of this house price prediction model's performance, let's create a small test dataset and see how accurately the model will be on the test dataset you create!

In the following, we have collected information on 8 candidate houses in Iowa, in terms of their living area size, finish quality (between 1 and 10), and the number of bedrooms and bathrooms. Please select 5 of them into your own test dataset!

Acuta Property Advisors	800 sqft, Quality: 2, 1 bds, 1 ba	<input checked="" type="checkbox"/>
Brick Lane Property	988 sqft, Quality: 5, 3 bds, 1 ba	<input type="checkbox"/>
Corsair Realty	1088 sqft, Quality: 8, 2 bds, 1 ba	<input type="checkbox"/>
Ames Home	1733 sqft, Quality: 3, 4 bds, 2 ba	<input type="checkbox"/>
Excelsior Lane Realty	1721 sqft, Quality: 6, 3 bds, 2 ba	<input checked="" type="checkbox"/>
Finders Property	1973 sqft, Quality: 8, 3 bds, 2 ba	<input checked="" type="checkbox"/>
Gnu property	2290 sqft, Quality: 3, 4 bds, 2 ba	<input type="checkbox"/>
Herringbone Real Estate	2144 sqft, Quality: 6, 4 bds, 2 ba	<input checked="" type="checkbox"/>
Ivy Home	2945 sqft, Quality: 10, 3 bds, 3 ba	<input checked="" type="checkbox"/>

[Next](#)

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

Review your test dataset

Your test dataset is composed of the following 5 houses:

Acuta Property Advisors	800 sqft, Quality: 2, 1 bds, 1 ba
Excelsior Lane Realty	1721 sqft, Quality: 6, 3 bds, 2 ba
Finders Property	1973 sqft, Quality: 8, 3 bds, 2 ba
Herringbone Real Estate	2144 sqft, Quality: 6, 4 bds, 2 ba
Ivy Home	2945 sqft, Quality: 10, 3 bds, 3 ba

Note that within your test dataset:

- 20.00% of houses has a size that is smaller than 1200 sqft
- 40.00% of houses has a size that is between 1200 and 2000 sqft
- 40.00% of houses has a size that is larger than 2000 sqft

Whenever you are ready, click the button below to start evaluating the house price prediction model.

[Test the model](#)

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

#### How well the model performs on your test dataset?

Before revealing the model's performance on your selected test dataset to you, let's first make a guess. As a hint, the mean absolute percentage error (MAPE) of the model was 15.93% on the **training dataset**.

What do you think would be the model's MAPE on your test dataset? Make a guess below.

MAPE on test dataset (your guess) =

%

Next

#### How well the model performs on your test dataset?

Before revealing the model's performance on your selected test dataset to you, let's first make a guess. As a hint, the mean absolute percentage error (MAPE) of the model was 15.93% on the **training dataset**.

What do you think would be the model's MAPE on your test dataset? Make a guess below.

MAPE on test dataset (your guess) = 16%  
 Actual MAPE = 25.11%  
You have underestimated the model's error rate on the selected test dataset by 9.11%.

Note that the model's MAPE on the selected test dataset is different from its MAPE on the training dataset. Why will this happen?

[Let's see why](#)

(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

House Price Prediction

Prediction Task (Phase 1): 1/10

Please review the information below and predict the sale price of the house.

House Detail			
1. Living Area in Sq. Ft: 2320	2. Garage in Sq. Ft: 672		
3. Bedroom: 2	4. Bathroom: 2		
5. Story: 2	6. House quality: 5		
7. Year Built: 1948	8. Year Sold: 2006		

Make your own prediction on the sale price of this house.

[Submit & View Model's Prediction](#)

House Price Prediction

Prediction Task (Phase 1): 1/10

House Detail			
1. Living Area in Sq. Ft: 2320	2. Garage in Sq. Ft: 672		
3. Bedroom: 2	4. Bathroom: 2		
5. Story: 2	6. House quality: 5		
7. Year Built: 1948	8. Year Sold: 2006		

Your Prediction

Model's Prediction

[View the ground truth](#)

House Price Prediction

Prediction Task (Phase 1): 1/10

House Detail			
1. Living Area in Sq. Ft: 2320	2. Garage in Sq. Ft: 672		
3. Bedroom: 2	4. Bathroom: 2		
5. Story: 2	6. House quality: 5		
7. Year Built: 1948	8. Year Sold: 2006		

Your Prediction

<der>

Model's Prediction

Actual Sale Price

[Next Task](#)

## 6 THE INTERFACE OF THE MID-POINT FEEDBACK PAGE

House Price Prediction

### Phase 1 Summary

Congratulations! You have just completed all 10 tasks in Phase 1!

Before moving on to Phase 2, let's take a quick look at that in Phase 1 tasks, how accurate your predictions of house price are, as well as how accurate the machine learning model is.

Given a prediction that is made by you (or the machine learning model), its difference to the actual sale price of the house is calculated as the absolute difference between the prediction and the actual price, divided by the actual price. The table below shows for each task in Phase 1, the difference between your prediction and the actual house price, as well as the difference between the machine learning model's prediction and the actual house price.

Task No.	Actual Sale Price	Your Prediction	Difference (%): You vs. Actual	Model's prediction	Difference (%): Model vs. Actual
1	169000	120000	28.99%	177941	5.29%
2	141500	160000	13.07%	145376	2.74%
3	84900	90000	6.01%	101389	19.42%
4	89500	156000	74.30%	158425	77.01%
5	175000	180000	2.86%	168474	3.73%
6	127000	140000	10.24%	134786	6.13%
7	168000	160000	4.76%	165346	1.58%
8	127500	120000	5.88%	144373	13.23%
9	128000	120000	6.25%	107102	16.33%
10	113000	105000	7.08%	122402	8.32%
Average			15.94%		15.38%

When you finish reviewing this information, click the button below to proceed to the phase 2.

[Go to Phase 2](#)

## 7 THE TASK INTERFACE OF PHASE 2

House Price Prediction

Prediction Task (Phase 2): 1/10

Please review the information below and predict the sale price of the house.

House Detail	
1. Living Area in Sq. Ft: 813	2. Garage in Sq. Ft: 270
3. Bedroom: 2	4. Bathroom: 1
5. Story: 1	6. House quality: 6
7. Year Built: 1930	8. Year Sold: 2006

Make your first prediction on the sale price of this house.

\$

House Price Prediction

Prediction Task (Phase 2): 1/10

Please review the information below and predict the sale price of the house.

House Detail	
1. Living Area in Sq. Ft: 813	2. Garage in Sq. Ft: 270
3. Bedroom: 2	4. Bathroom: 1
5. Story: 1	6. House quality: 6
7. Year Built: 1930	8. Year Sold: 2006

 Your Initial Prediction  
90000

 Model's Prediction  
133699

Please drag the slider or input to the textbox to make Your Final Prediction

\$