

Intro 1

People often rely on machine learning model outputs to make decisions.

Many factors can contribute to a machine learning model's output. For example, the output of a rain-predicting model can rely on factors such as the current temperature and wind speed.

Computer scientists refer to these factors as **model explanations**.

We will teach you how to interpret these explanations and ask you questions about them.

Intro 2

Someone designed a machine learning model to predict whether it is a good idea to put on a coat or not.

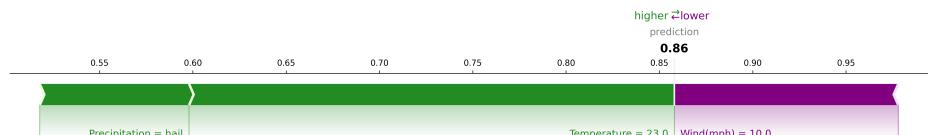
It calculates the probability that you should put on a coat using

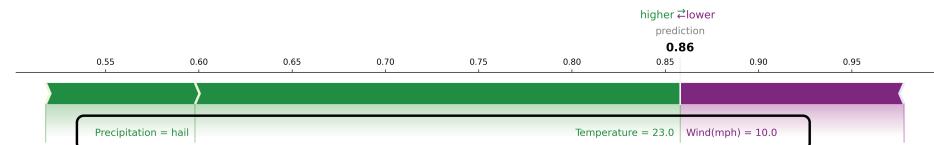
the current temperature, wind speed, and precipitation.

If that probability is greater than or equal to 0.5, then the model will recommend that you put on a coat. If the probability is less than 0.5, then the model will recommend that you do NOT put on a coat.

Intro 3

Below, you can see a visual explanation for one instance of the model prediction, based on some input values for the three factors the model considers (temperature, wind speed, and precipitation).





Let's take a closer look at this visual explanation.

Intro 4

Under the green and purple bars, you can see the factors that the model uses to make predictions.

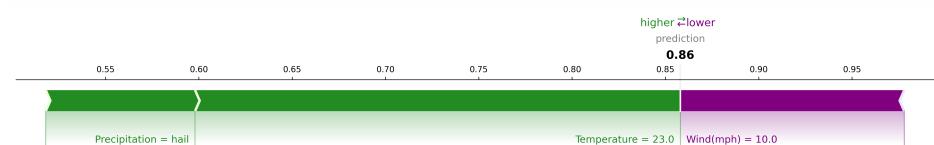
This model takes three factors into account when making predictions: temperature, wind, and precipitation.

These factors can take inputs that are numerical (e.g., 30, 0) or categorical (e.g., rain, snow).

Intro 4.5

You can also mouse over the new "See table view" button below to see another view of the features. Give it a try!

A factor's exact value will be shown in the **Value** column in the table on the right.



[See table view](#)

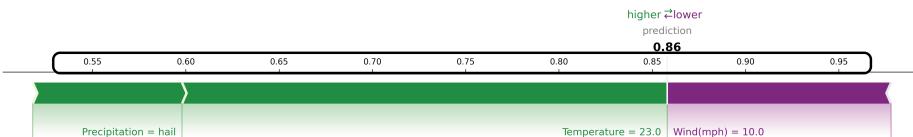
Feature	Value
Temperature	23
Wind	10
Precipitation=hail	True

See table view

Intro 5

The line above the bar shows the probability value generated by the model.

This probability describes whether it is a good idea to put on a coat or not (probability ≥ 0.5 , good idea to put on a coat; probability < 0.5 , NOT a good idea).



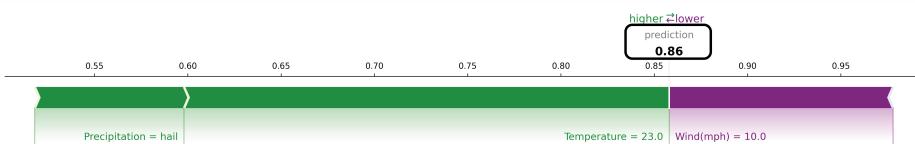
Intro 6

You can put different values of temperature, wind, and precipitation into your model to generate a **prediction**. This generated prediction probability is also labeled on the graph.

If the prediction is **greater than or equal** to 0.5, the model will return 'YES', suggesting that you should wear a coat. If the prediction is **less than** 0.5, the model will return 'NO', suggesting that you do not wear a coat.

The visualization shows how, starting from the **base value**, each input values of temperature, wind, and precipitation can have a

positive (green) contribution, pushing the prediction toward 'YES', or a negative (purple) contribution, pushing the prediction toward 'NO'.

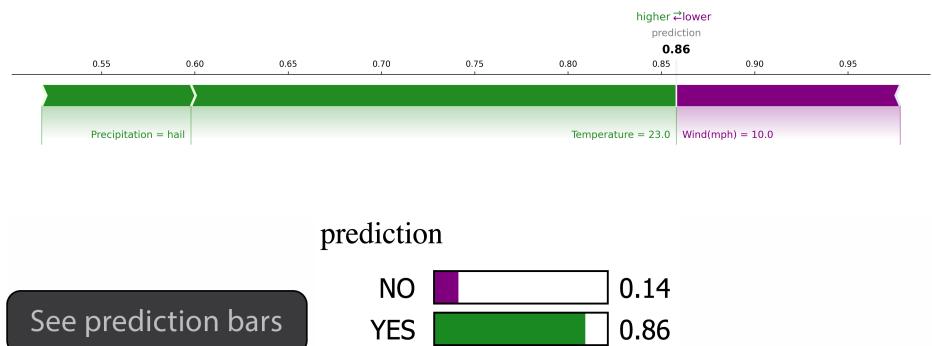


Feature	Value
Temperature	23
Wind	10
Precipitation=hail	True

See table view

You can also mouse over the new "see prediction bars" button below to see another prediction visualization. Give it a try!

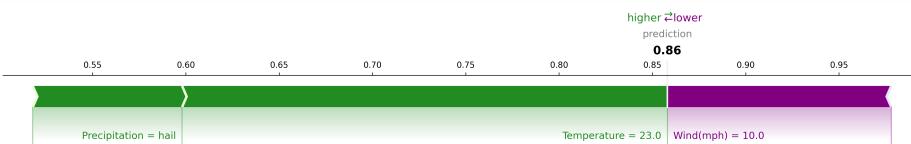
Here you can see a green bar giving the model's prediction model for 'YES', and a purple bar giving the model's prediction model for 'NO'.



Intro 6.5

Intro 7

From now on, you will be able to mouse over both buttons. Give it a try!



[See table view](#)

[See prediction bars](#)

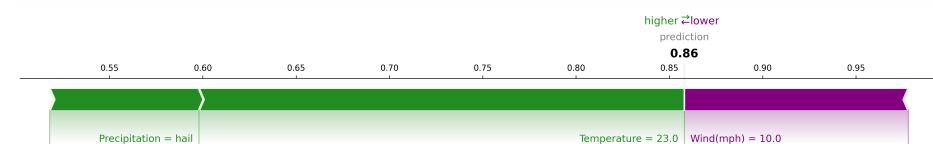
prediction



Intro Test 1

In the example below, what will the model predict?

- YES, you should wear a coat
- NO, do not wear a coat



[See prediction bars](#)

[See table view](#)

Feature	Value
Temperature	23
Wind	10
Precipitation	hail

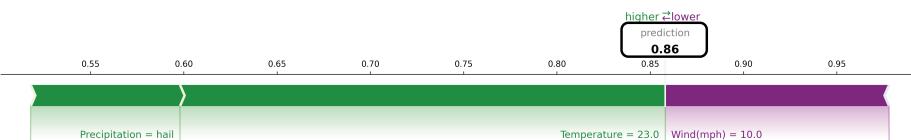
Correct. In this case, the model prediction is 0.86, which is larger than 0.5, so the model will return YES. You can see this value above the green and purple bars, or by mousing over the "See prediction bars" button.

Feature	Value
Temperature	23
Wind	10
Precipitation	hail

Not quite. In this case, the model prediction is 0.86, which is larger than 0.5, so the model will return YES.

Intro Test 2

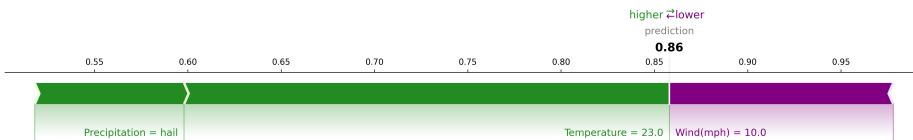
As another review, by looking at the explanation image, please select the value for **precipitation** input into the model:



[See prediction bars](#)

[See table view](#)

- sleet
- snow
- hail
- rain
- none



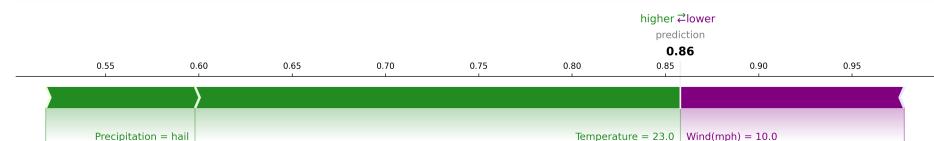
[See prediction bars](#)

[See table view](#)

	Feature	Value
prediction	Temperature	23
	Wind	10
NO	0.14	
YES	0.86	hail

Correct - the value is printed next to the word **Precipitation** under the red and blue bars, or next to the word "Precipitation" in the table you see when you mouse over "See table view". This value is **hail**.

Not quite - the value is printed next to the word **Precipitation** under the red and blue bars, or next to the word "Precipitation" in the table you see when you mouse over "See table view". This value is **hail**.



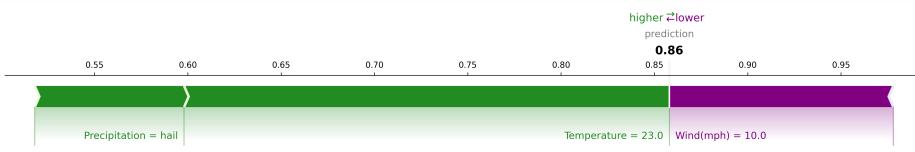
[See prediction bars](#)

[See table view](#)

	Feature	Value
prediction	Temperature	23
	Wind	10
NO	0.14	
YES	0.86	hail

By looking at the explanation image, please select the value for **wind speed** input into the model:

- 20 mph
- 0 mph
- 10 mph
- 5 mph
- 15 mph



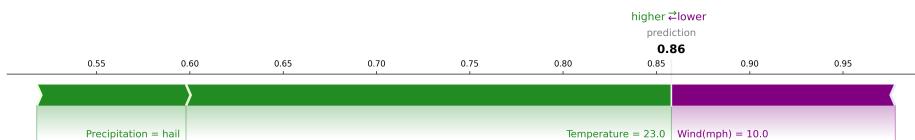
[See prediction bars](#)

[See table view](#)

Feature	Value
Temperature	23
Wind	10
Precipitation	hail

Correct – the value is printed next to the word **Wind(mph)** under the red and blue bars, or next to the word "Wind" in the table you see when you mouse over "See table view". This value is **10 mph**.

Not quite – the value is printed next to the word **Wind(mph)** under the red and blue bars, or next to the word "Wind" in the table you see when you mouse over "See table view". This value is **10 mph**.

**See prediction bars****See table view**

	Feature	Value
prediction	Temperature	23
	Wind	10
	Precipitation	hail
	NO	0.14
YES	0.86	

The wind factor, in this example, with input value of 10 mph, pushes the model prediction *lower*. This means the current value of Wind(mph) is pushing the model toward predicting 'NO'.

Factors that push the model toward predicting 'NO' are always colored **purple** and point to the *left*.

If the final prediction is pushed below 0.5, the model will return 'NO' (do not wear a coat).

**See prediction bars****See table view**

	Feature	Value
prediction	Temperature	23
	Wind	10
	Precipitation	hail
	NO	0.14
YES	0.86	

Intro 8

Each input value of temperature, wind, and precipitation can push the model's **prediction** to be higher or lower.

If the final prediction is pushed to 0.5 or above, the model will return 'YES' (wear a coat).

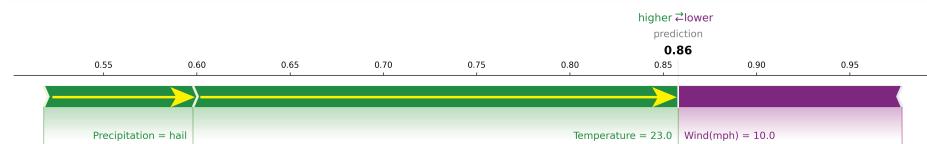
You can also see the pushing direction of factor by looking at the table view.

Rows containing factors that push the model toward predicting 'NO' are always colored **purple**.

Intro 9

The temperature and precipitation factors, with input value of '23' and 'hail', push the **prediction** higher. This means the current values of Temperature and Precipitation are pushing the model toward predicting 'YES'.

Factors that push the model toward predicting 'YES' are always colored **green** and always point to the *right*.



[See prediction bars](#)

[See table view](#)

	Feature	Value
prediction	Temperature	23
NO	Wind	10
YES	Precipitation	hail

In the table view, rows containing factors that push the model toward predicting 'YES' are always colored **green**.

	Feature	Value
prediction	Temperature	23
NO	Wind	10
YES	Precipitation	hail

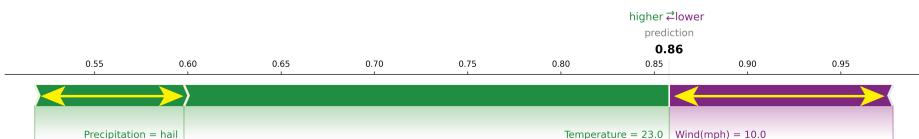
Intro 10

The **length** of a bar and the value inside it indicate the predictive power of a factor.

The wind factor has a **greater** predictive power compared to the precipitation factor. This means that the wind factor influences the model prediction more than the precipitation factor.

You can also see the power of a factor by looking at the table view. The features are sorted by impact.

Because the wind factor has more influence than the precipitation factor, you can see the row containing Wind above the row containing Precipitation.



See prediction bars

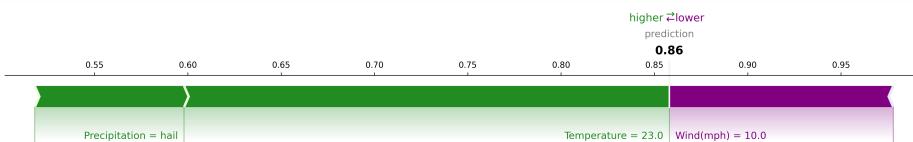
See table view

Intro Test 3

As a review, by looking at the explanation image, which factor(s)

are pushing the model toward predicting 'YES'?

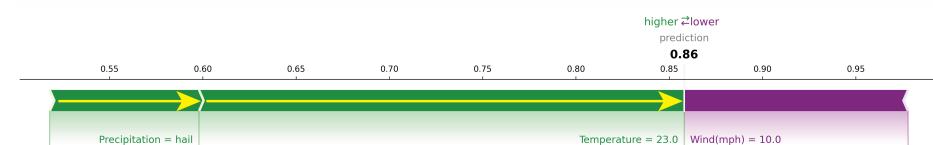
- Temperature
- Wind
- Precipitation



[See prediction bars](#)

[See table view](#)

	Feature	Value
prediction	Temperature	23
	Wind	10
NO	Precipitation=hail	True
YES		0.86

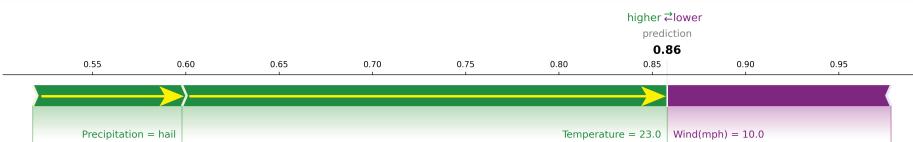


[See prediction bars](#)

[See table view](#)

	Feature	Value
prediction	Temperature	23
	Wind	10
NO	Precipitation	hail
YES		0.86

Not quite – In this case, the bars for temperature and precipitation are **green** and pointing to the **right**, so the values of these factors are pushing the prediction *higher* and pushing the model toward predicting 'YES'.



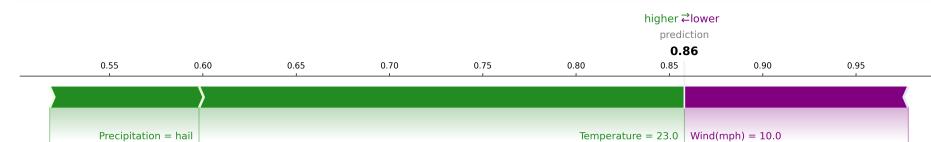
[See prediction bars](#)

[See table view](#)

	Feature	Value
prediction	Temperature	23
NO	Wind	10
YES	Precipitation	hail

By looking at the explanation image, which factor(s) are pushing the model toward predicting 'NO'?

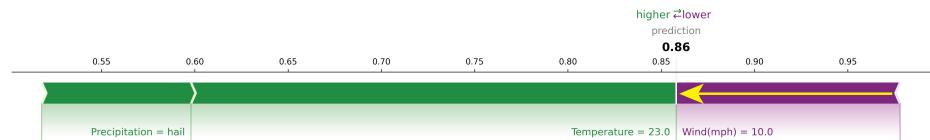
- Temperature
- Wind
- Precipitation

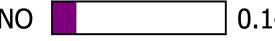


[See prediction bars](#)

[See table view](#)

	Feature	Value
prediction	Temperature	23
	Wind	10
	Precipitation	hail
NO		0.14
YES		0.86

[See prediction bars](#)[See table view](#)

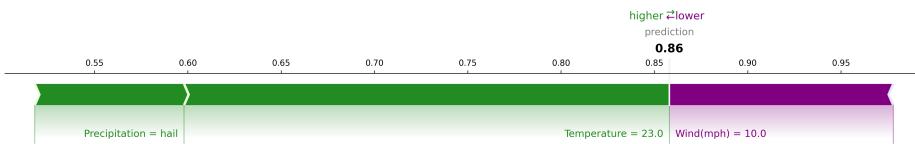
	Feature	Value
prediction	Temperature	23
	Wind	10
	Precipitation	hail
NO		0.14
YES		0.86

Correct – In this case, the bar for Wind(mph) is **purple** and pointing to the **left**, so the value of this factor is pushing the prediction *lower* and pushing the model toward predicting 'NO'.

Not quite – In this case, the bar for Wind(mph) is **purple** and pointing to the **left**, so the value of this factor is pushing the prediction *lower* and pushing the model toward predicting 'NO'.

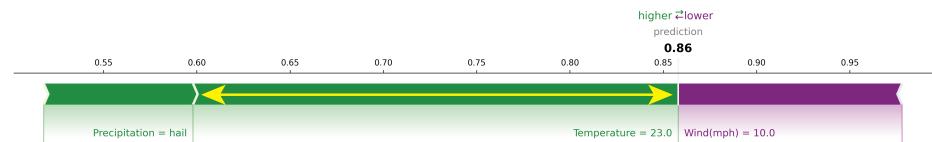
Which factor has the greatest predictive power?

- Temperature
- Wind
- Precipitation

**See prediction bars****See table view**

	Feature	Value
prediction	Temperature	23
NO	Wind	10
YES	Precipitation	hail

Not quite – In this case, Temperature has the **longest** bar and is the top row of the table, so Temperature has the greatest predictive power.

**See prediction bars****See table view**

	Feature	Value
prediction	Temperature	23
NO	Wind	10
YES	Precipitation	hail

Correct – In this case, Temperature has the **longest** bar and is the top row of the table, so Temperature has the greatest predictive power.

Intro Test 4

As a final review, what does the following model recommend you do?

- YES, you should wear a coat
- NO, do not wear a coat



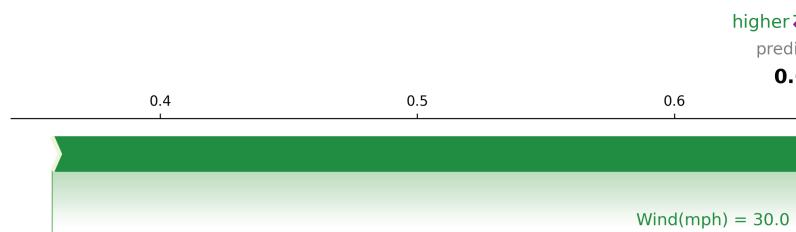
See prediction bars

See table view

	Feature	Value
prediction	Wind	30
NO	Temperature	70
YES	Precipitation	none

Correct. In this case, the model prediction is 0.65, which is greater than 0.5, so the model will return 'YES'. You can see this number above the bars, and next to the 'YES' when you mouse over "See prediction bars".

Incorrect. In this case, the model prediction is 0.65, which is greater than 0.5, so the model will return 'YES'. You can see this number above the bars, and next to the 'YES' when you mouse over "See prediction bars".



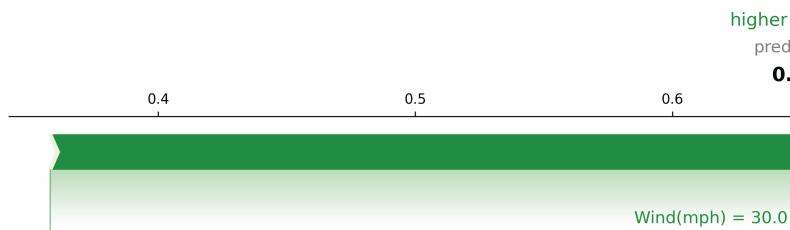
[See prediction bars](#)

[See table view](#)

Feature	Value
Wind	30
Temperature	70
Precipitation	none

By looking at the explanation image, please select the value for **temperature** input into the model:

- 84
- 70
- 61
- 56
- 37



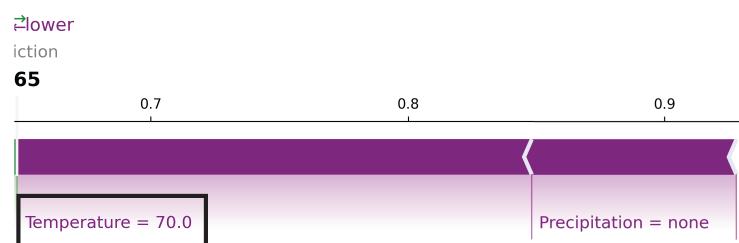
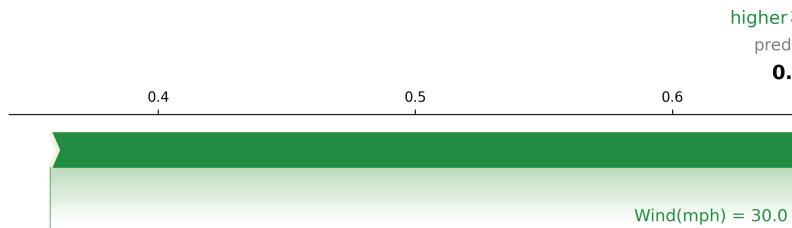
See prediction bars

See table view

	Feature	Value
prediction	Wind	30
NO	Temperature	70
YES	Precipitation	none

Correct – the value is next to the word **Temperature** under the green and purple bars, and in the bar graph you see when you mouse over "See table view". This value is **70**.

Incorrect – the value is next to the word **Temperature** under the green and purple bars and in the table you see when you mouse over "See table view". This value is **70**.



[See prediction bars](#)

[See table view](#)

	Feature	Value
prediction	Wind	30
NO	Temperature	70
YES	Precipitation	none

By looking at the explanation image, which factor(s) are pushing the model toward predicting 'NO'?

- Temperature
- Wind
- Precipitation

[See prediction bars](#)
[See table view](#)

	Feature	Value
prediction	Wind	30
NO	Temperature	70
YES	Precipitation	none

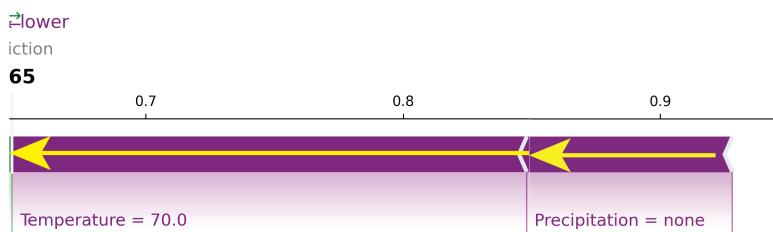
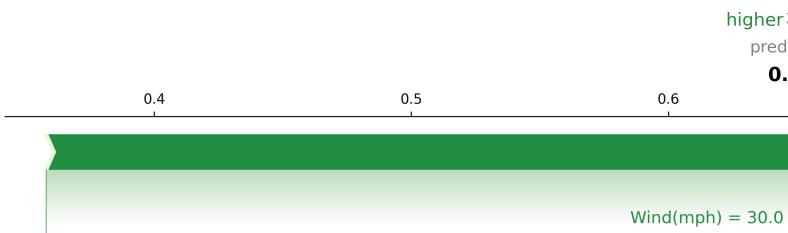


Correct – In this case, the bars for Temperature and Precipitation are **purple** and their bars are pointing to the **left**, so the values of these factors are pushing the prediction *lower* and pushing the model toward predicting 'NO'.

Not quite – In this case, the bars for Temperature and

Precipitation are **purple** and their bars are pointing to the **left**, so the values of these factors are pushing the prediction *lower* and pushing the model toward predicting 'NO'.

Feature	Value
Wind	30
Temperature	70
Precipitation	none

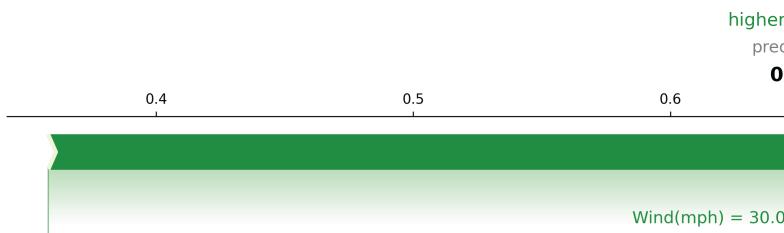


See prediction bars

See table view

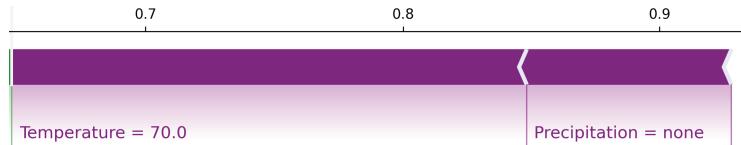
Which factor has the greatest predictive power?

- Temperature
- Wind
- Precipitation



lower
iction

65



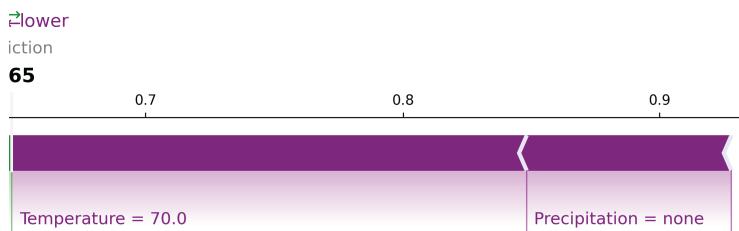
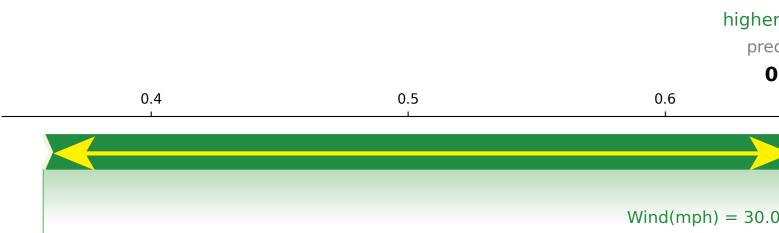
See prediction bars

See table view

	Feature	Value
prediction	Wind	30
NO	Temperature	70
YES	Precipitation	none

Correct – In this case, Wind(mph) has the **longest** bar, and is in the top row of the table you see when you mouse over "See table view", so Wind has the greatest predictive power.

Not quite – In this case, Wind(mph) has the longest bar, and is in the top row of the table you see when you mouse over "See table view", so Wind has the greatest predictive power.

**See prediction bars****See table view**

	Feature	Value
prediction	Wind	30
NO	Temperature	70
YES	Precipitation	none

Intro Main

We have another machine learning model that makes predictions to approve or deny a loan based on a set of factors related to the loan applicant.

The model is trained to predict a person's likely income using real data from 26,000 people, and uses this prediction to decide whether a person is likely to be able to pay back a loan. If the person is likely, the model outputs 'YES', they should be given a loan. If the person is not likely, the model outputs 'NO', they should not be given a loan.

The model generates a prediction based on each set of input values. If the predicted value is greater than or equal to 0.5, then

the model will approve the loan. If the predicted value is less than 0.5, the model will deny the loan.

Six people applied to the loan. We input their corresponding values for each factor into the model.

We will show you six predictions the models generated for each of the six loan applicants.

Keep in mind that all six predictions were made by the **same** model.

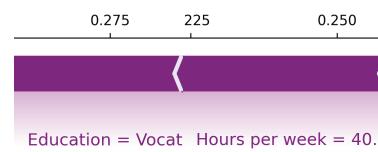
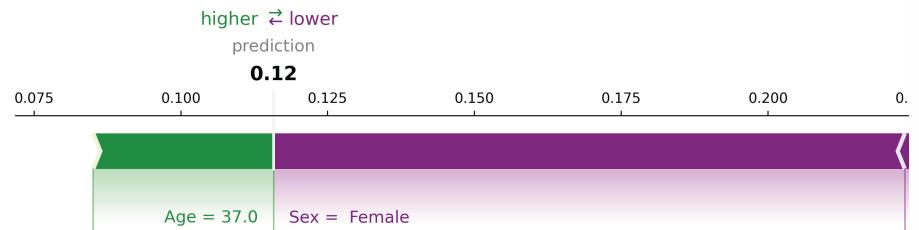
Woman 1

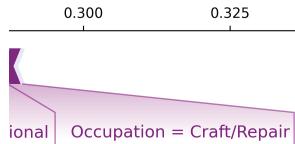
Below you will find the information of Applicant X.

You can see that the model made a prediction of whether to approve or deny a loan from this applicant based on five factors. The explanation is below.

Look at the explanation, and answer the questions that follow.

Remember that if the model's prediction probability (predicted value) for 'YES' is greater than or equal to 0.5, the model will return 'YES' (approve the loan). If it is less than 0.5, the model will return 'NO' (deny the loan).





[See prediction bars](#)

[See table view](#)

prediction



NO **0.88**

YES **0.12**

Feature	Value
Sex	Female
Hours	40
Age	37
Education	Vocational
Occupation	Craft/Repair

Will this model approve the loan for this person?

YES

NO

What feature was had the most predictive power for this decision?

- Education
- Hours Worked Per Week
- Age
- Sex
- Occupation

Which factor(s) are pushing the model toward predicting 'NO'?

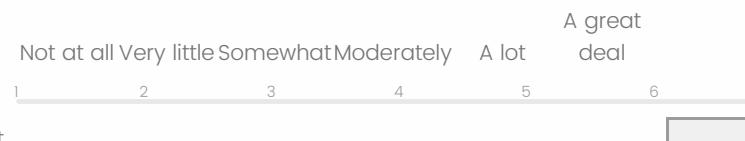
- Education
- Hours Worked Per Week
- Age
- Sex
- Occupation
- None of these

Which factor(s) are pushing the model toward predicting 'YES'?

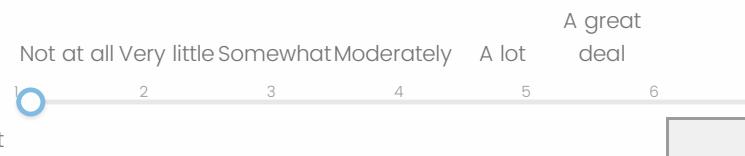
- Education
- Hours Worked Per Week

- Age
- Sex
- Occupation
- None of these

On a scale from 1 to 6, how much do you trust the model to approve or deny a loan for **you**?

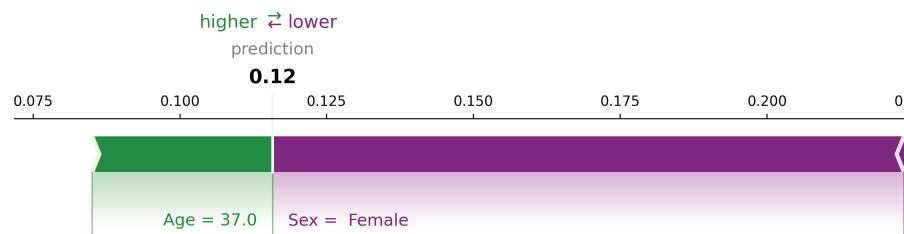


On a scale from 1 to 6, how much do you trust the model to approve or deny a loan for **other people in general**?

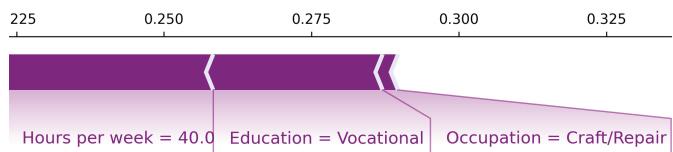


Please indicate whether you agree with the below statements.

- | | Agree |
|---|-----------------------|
| This model uses all of the features that it should use when making this decision. | <input type="radio"/> |
| This model does not use any unnecessary features when making this decision. | <input type="radio"/> |
| I trust the data this model was trained on. | <input type="radio"/> |
| Computer models can be trusted to make human decisions. | <input type="radio"/> |
| This model is accurate. | <input type="radio"/> |
| This model is fair. | <input type="radio"/> |
| This model would probably give me a loan because I am similar to the person described in this question. | <input type="radio"/> |
| This model would probably give me a loan because I am different from the person described in this question. | <input type="radio"/> |
| This model would probably give me a loan because of previous decisions it has made. | <input type="radio"/> |
| This model probably would not give me a loan, and this would be the correct decision. | <input type="radio"/> |



When answering the previous questions about the given explanation, which design aspects of the visualizations did you find **most** useful?



When answering the previous questions about the given explanation, which design aspects of the visualizations did you find **least** useful?

See prediction bars

See table view

prediction

NO 0.88

YES 0.12

Feature	Value
Sex	Female
Hours	40
Age	37
Education	Vocational
Occupation	Craft/Repair

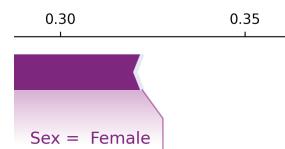
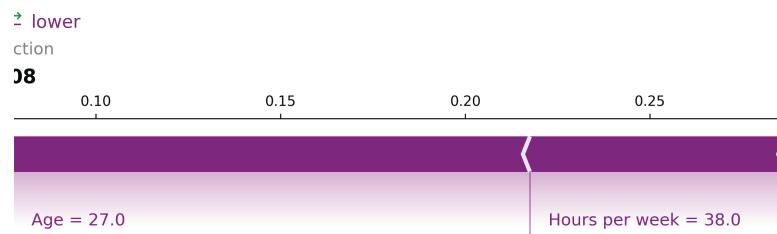
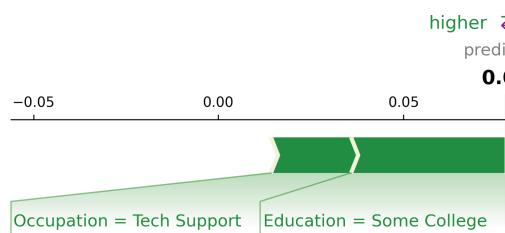
Woman 2

Below you will find the information of Applicant R.

You can see that the model made a prediction of whether to approve or deny a loan from this applicant based on five factors. The explanation is below.

Look at the explanation, and answer the questions that follow.

Remember that if the model's prediction probability (predicted value) for 'YES' is greater than or equal to 0.5, the model will return 'YES' (approve the loan). If it is less than 0.5, the model will return 'NO' (deny the loan).



[See prediction bars](#) [See table view](#)

Feature	Value
Age	27
Hours	38
Education	Some College
Sex	Female
Occupation	Tech Support

Will this model approve the loan for this person?

- YES
- NO

Which feature was had the most predictive power for this decision?

- Education
- Hours Worked Per Week
- Age
- Sex
- Occupation

Which factor(s) are pushing the model toward predicting 'NO'?

- Education
- Hours Worked Per Week
- Age
- Sex
- Occupation
- None of these

Which factor(s) are pushing the model toward predicting 'YES'?

- Education
- Hours Worked Per Week
- Age
- Sex
- Occupation
- None of these

On a scale from 1 to 6, how much do you trust the model to approve or deny a loan for **you**?



On a scale from 1 to 6, how much do you trust the model to approve or deny a loan for **other people in general**?



Please indicate whether you agree with the below statements.

Agree

-
-
-
-
-
-
-
-
-
-

This model uses all of the features that it should use when making this decision.

This model does not use any unnecessary features when making this decision.

I trust the data this model was trained on.

Computer models can be trusted to make human decisions.

This model is accurate.

This model is fair.

This model would probably give me a loan because I am similar to the person described in this question.

This model would probably give me a loan because I am different from the person described in this question.

This model would probably give me a loan because of previous decisions it has made.

This model probably would not give me a loan, and this would be the correct decision.

Woman 3

Below you will find the information of Applicant S.

You can see that the model made a prediction of whether to approve or deny a loan from this applicant based on five factors. The explanation is below.

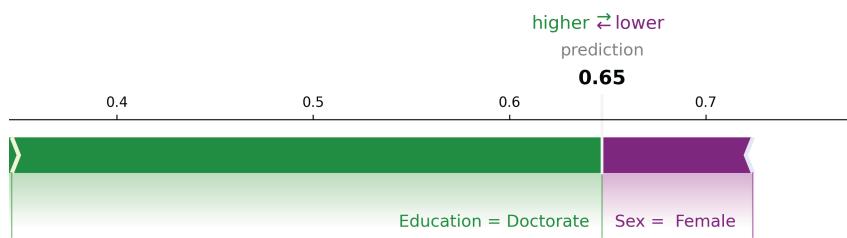
Look at the explanation, and answer the questions that follow.

Remember that if the model's prediction probability (predicted value) for 'YES' is greater than or equal to 0.5, the model will return 'YES' (approve the loan). If it is less than 0.5, the model will return 'NO' (deny the loan).



Will this model approve the loan for this person?

- YES
- NO



Which feature had the most predictive power for this decision?

- Education
- Hours Worked Per Week
- Age
- Sex
- Occupation

See prediction bars

See table view

prediction

NO  0.35

YES  0.65

Feature	Value
Education	Doctorate
Age	51
Sex	Female
Occupation	Exec. Managerial
Hours	45

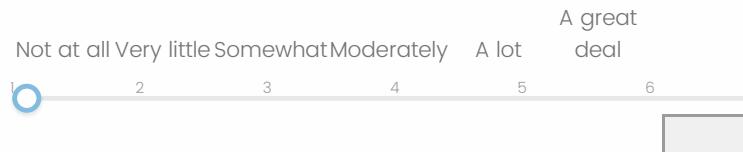
Which factor(s) are pushing the model toward predicting 'NO'?

- Education
- Hours Worked Per Week
- Age
- Sex
- Occupation
- None of these

Which factor(s) are pushing the model toward predicting 'YES'?

- Education
- Hours Worked Per Week
- Age
- Sex
- Occupation
- None of these

On a scale from 1 to 6, how much do you trust the model to approve or deny a loan for **you**?



On a scale from 1 to 6, how much do you trust the model to approve or deny a loan for **other people in general**?

A great



Please indicate whether you agree with the below statements.

- | | Agree |
|---|-----------------------|
| This model uses all of the features that it should use when making this decision. | <input type="radio"/> |
| This model does not use any unnecessary features when making this decision. | <input type="radio"/> |
| I trust the data this model was trained on. | <input type="radio"/> |
| Computer models can be trusted to make human decisions. | <input type="radio"/> |
| This model is accurate. | <input type="radio"/> |
| This model is fair. | <input type="radio"/> |
| This model would probably give me a loan because I am similar to the person described in this question. | <input type="radio"/> |
| This model would probably give me a loan because I am different from the person described in this question. | <input type="radio"/> |
| This model would probably give me a loan because of previous decisions it has made. | <input type="radio"/> |
| This model probably would not give me a loan, and this would be the correct decision. | <input type="radio"/> |

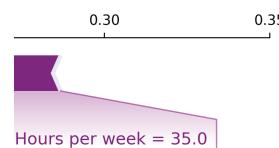
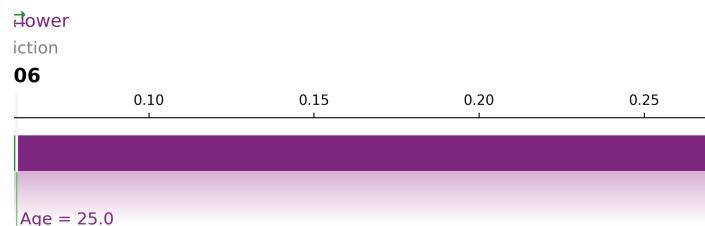
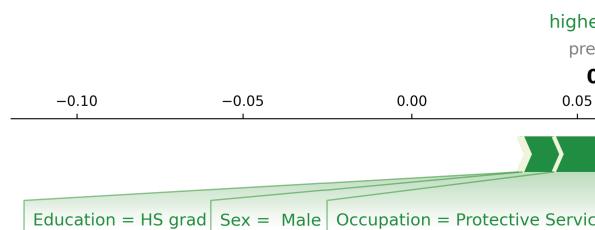
Man 1

Below you will find the information of Applicant N.

You can see that the model made a prediction of whether to approve or deny a loan from this applicant based on five factors. The explanation is below.

Look at the explanation, and answer the questions that follow.

Remember that if the model's prediction probability (predicted value) for 'YES' is greater than or equal to 0.5, the model will return 'YES' (approve the loan). If it is less than 0.5, the model will return 'NO' (deny the loan).



[See prediction bars](#)

[See table view](#)



- Education
- Hours Worked Per Week
- Age
- Sex
- Occupation
- None of these

Will this model approve the loan for this person?

- YES
- NO

Which feature had the most predictive power for this decision?

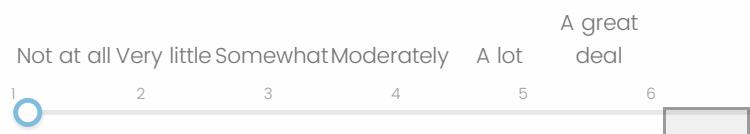
- Education
- Hours Worked Per Week
- Age
- Sex
- Occupation

Which factor(s) are pushing the model toward predicting 'NO'?

Which factor(s) are pushing the model toward predicting 'YES'?

- Education
- Hours Worked Per Week
- Age
- Sex
- Occupation
- None of these

On a scale from 1 to 6, how much do you trust the model to approve or deny a loan for **you**?



Level of Trust

On a scale from 1 to 6, how much do you trust the model to approve or deny a loan for **other people in general?**



Please indicate whether you agree with the below statements.

This model uses all of the features that it should use when making this decision.

This model does not use any unnecessary features when making this decision.

I trust the data this model was trained on.

Computer models can be trusted to make human decisions.

This model is accurate.

This model is fair.

This model would probably give me a loan because I am similar to the person described in this question.

Agree

-
-
-
-
-
-

This model would probably give me a loan because I am different from the person described in this question.

This model would probably give me a loan because of previous decisions it has made.

This model probably would not give me a loan, and this would be the correct decision.

Man 2

Below you will find the information of Applicant P.

You can see that the model made a prediction of whether to approve or deny a loan from this applicant based on five factors.

The explanation is below. Look at the explanation, and answer the questions that follow.

Remember that if the model's prediction probability (predicted value) for 'YES' is greater than or equal to 0.5, the model will return 'YES' (approve the loan). If it is less than 0.5, the model will return 'NO' (deny the loan).

[See prediction bars](#)[See table view](#)

Feature	Value
Education	Bachelors
Sex	Male
Hours	50
Age	38
Occupation	Sales

Will this model approve the loan for this person?

- YES
- NO

Which feature had the most predictive power for this decision?

- Education
- Hours Worked Per Week
- Age
- Sex
- Occupation

Which factor(s) are pushing the model toward predicting 'NO'?

- Education
- Hours Worked Per Week
- Age
- Sex
- Occupation
- None of these



On a scale from 1 to 6, how much do you trust the model to approve or deny a loan for **other people in general**?

Which factor(s) are pushing the model toward predicting 'YES'?

- Education
- Hours Worked Per Week
- Age
- Sex
- Occupation
- None of these



On a scale from 1 to 6, how much do you trust the model to approve or deny a loan for **you**?

Not at all Very little Somewhat Moderately A lot A great deal

Please indicate whether you agree with the below statements.

- | | Agree |
|---|-----------------------|
| This model uses all of the features that it should use when making this decision. | <input type="radio"/> |
| This model does not use any unnecessary features when making this decision. | <input type="radio"/> |
| I trust the data this model was trained on. | <input type="radio"/> |
| Computer models can be trusted to make human decisions. | <input type="radio"/> |
| This model is accurate. | <input type="radio"/> |
| This model is fair. | <input type="radio"/> |

This model would probably give me a loan because I am similar to the person described in this question.

This model would probably give me a loan because I am different from the person described in this question.

This model would probably give me a loan because of previous decisions it has made.

This model probably would not give me a loan, and this would be the correct decision.

See prediction bars

See table view

	Feature	Value
	Education	10th
	Sex	Male
	Hours	48
	Occupation	Transport/Moving
prediction		
NO		0.86
YES		0.14
	Age	36

Man 3

Below you will find the information of Applicant K.

You can see that the model made a prediction of whether to approve or deny a loan from this applicant based on five factors. The explanation is below.

Look at the explanation, and answer the questions that follow.

Remember that if the model's prediction probability (predicted value) for 'YES' is greater than or equal to 0.5, the model will return 'YES' (approve the loan). If it is less than 0.5, the model will return 'NO' (deny the loan).

Will this model approve the loan for this person?

- YES
- NO

What feature had the most predictive power for this decision?

- Education
- Hours Worked Per Week
- Age
- Sex
- Occupation

Which factor(s) are pushing the model toward predicting 'NO'?

- Education
- Hours Worked Per Week
- Age
- Sex
- Occupation
- None of these

Not at all Very little Somewhat Moderately A lot deal



Level of Trust

Which factor(s) are pushing the model toward predicting 'YES'?

- Education
- Hours Worked Per Week
- Age
- Sex
- Occupation
- None of these

Not at all Very little Somewhat Moderately A lot deal



Level of Trust

On a scale from 1 to 6, how much do you trust the model to approve or deny a loan for **you**?

A great

On a scale from 1 to 6, how much do you trust the model to approve or deny a loan for **other people in general**?

Please indicate whether you agree with the below statements.

Agree

This model uses all of the features that it should use when making this decision.

This model does not use any unnecessary features when making this decision.

I trust the data this model was trained on.

Computer models can be trusted to make human decisions.

This model is accurate.

This model is fair.

This model would probably give me a loan because I am similar to the person described in this question.

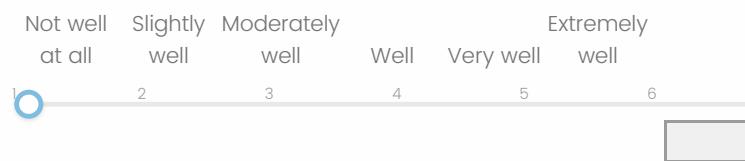
This model would probably give me a loan because I am different from the person described in this question.

This model would probably give me a loan because of previous decisions it has made.

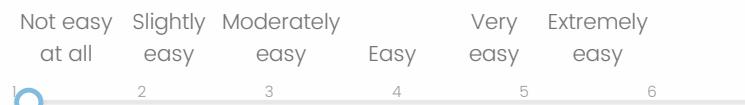
This model probably would not give me a loan, and this would be the correct decision.

Perception of understanding

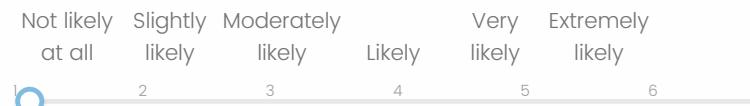
How well did you understand the way this model makes decisions?



How easy was it for you to understand the model output?



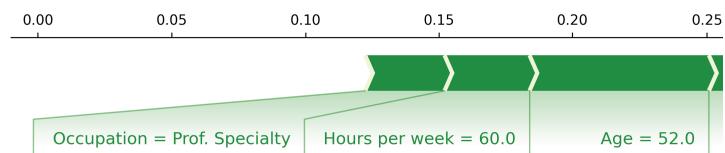
How likely would you use this visualization to explain models to other people?



Fairness

Below are two explanations for predictions made by the same loan approval machine learning model you have been seeing, for two people with almost identical features.

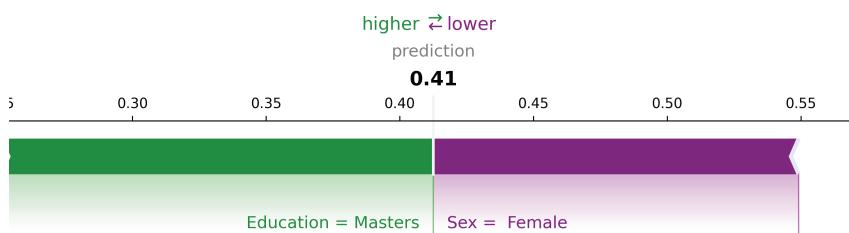
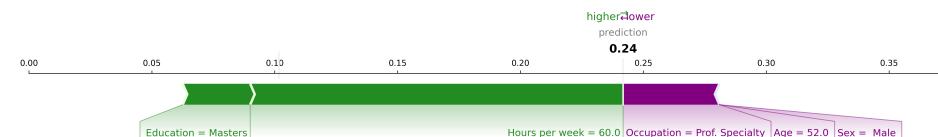
Remember that if the model's prediction probability (Predicted Value) for 'YES' is greater than or equal to 0.5, the model will return 'YES' (approve the loan). If it is less than 0.5, the model will return 'NO' (deny the loan).



Feature	Value
Education	Masters
Sex	Female
Age	52
Hours	60
Occupation	Prof. Specialty

prediction

NO 0.59
YES 0.41

**See prediction bars****See table view****See prediction bars****See table view**

Feature	Value
Education	Masters
Sex	Male
Age	52
Hours	60
Occupation	Prof. Specialty

prediction

NO 0.26
YES 0.74

Will this model approve the loan for **Person A**?

- YES
- NO



Will this model approve the loan for **Person B**?

- YES
- NO

On a scale from 1 to 6, how much do you trust the model to approve or deny a loan for **you**?

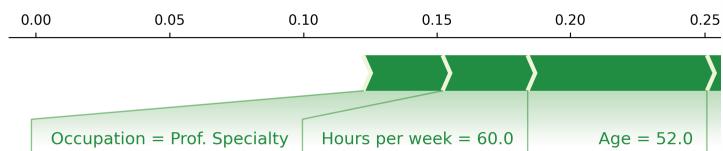


On a scale from 1 to 6, how much do you trust the model to approve or deny a loan for **other people in general**?

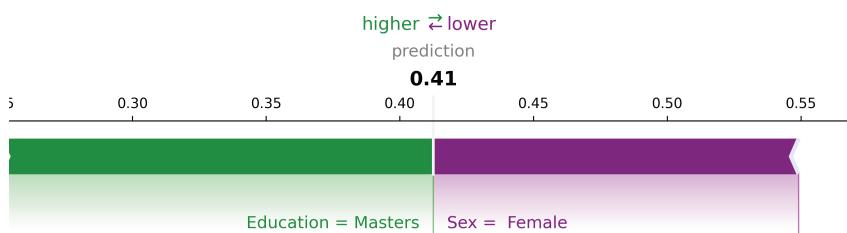
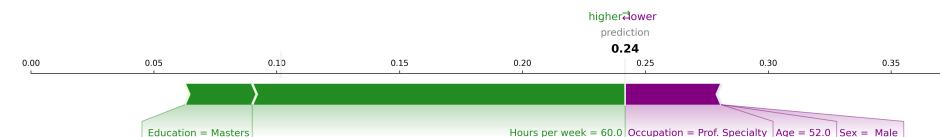
Please indicate whether you agree with the below statements.

Agree

- This model uses all of the features that it should use when making this decision.
- This model does not use any unnecessary features when making this decision.
- I trust the data this model was trained on.
- Computer models can be trusted to make human decisions.
- This model is accurate.
- This model is fair.
- This model would probably give me a loan because I am similar to a person described in this question.
- This model would probably give me a loan because I am different from a person described in this question.
- This model would probably give me a loan because of previous decisions it has made.
- This model probably would not give me a loan, and this would be the correct decision.

Fairness General

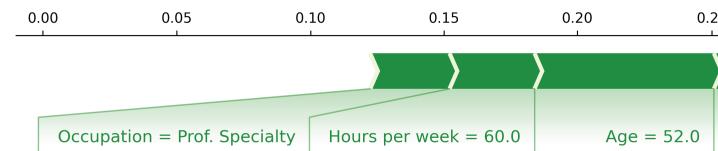
Feature	Value
Education	Masters
Sex	Female
Age	52
Hours	60
Occupation	Prof. Specialty

**See prediction bars****See table view****See prediction bars****See table view**

Feature	Value
Education	Masters
Sex	Male
Age	52
Hours	60
Occupation	Prof. Specialty

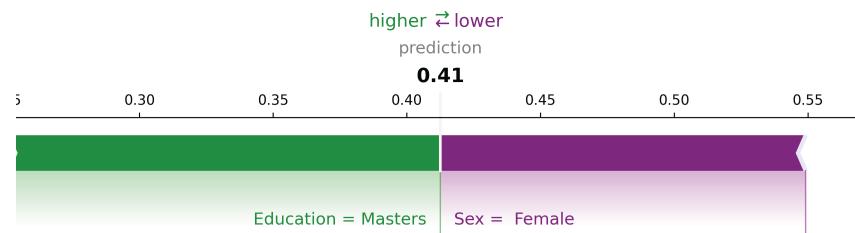
Do you think this model includes potentially discriminating factors?

- YES
- NO



If yes, which ones?

- Age
- Hours Per Week
- Education
- Occupation
- Sex

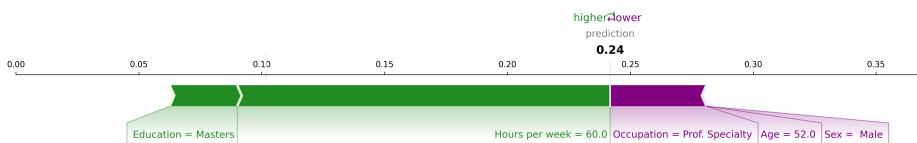


Block 38

See prediction bars

See table view

	Feature	Value
	Education	Masters
	Sex	Female
	Age	52
prediction	Hours	60
NO	Occupation	Prof. Specialty
YES		

**See prediction bars****See table view****prediction**

NO		0.26
YES		0.74

Feature	Value
Education	Masters
Sex	Male
Age	52
Hours	60
Occupation	Prof. Specialty

When answering the previous questions about fairness, which design aspects of the given visualizations did you find **most** useful?

When answering the previous questions about fairness, which design aspects of the given visualizations did you find **least** useful?

Demographics

What is your age? Please enter a number.

What is your gender?

- Man/Male (Cis or Trans)
- Woman/Female (Cis or Trans)
- Non-binary
- My Gender is Not Listed Above: (Open Text Box)
- Unsure/Questioning
- Prefer Not to Answer

What is your race/ethnicity?

- White
- Black/African American
- Hispanic/Latinx
- Asian
- Native American
- Hawaiian/Pacific Islander
- Other

How much is your yearly income?

- \$0 - \$49,999
- \$50,000 - \$99,999
- \$100,000+
- Other

What is the highest level of school you have completed or the highest degree you have received?

- Less than high school degree
- High school graduate (high school diploma or equivalent including GED)
- Some college but no degree
- Associate degree in college (2-year)
- Bachelor's degree in college (4-year)
- Master's degree
- Professional degree (JD, MD, PhD)
- Prefer not to answer

What is your familiarity with machine learning models?

- No familiarity
- Beginner
- Intermediate
- Expert

Feedback

Please give any feedback or suggestions you may have about this survey

Powered by Qualtrics