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.

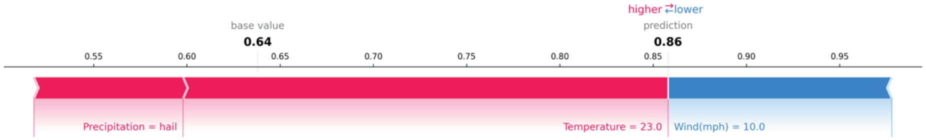
It calculates the probability that you should put on a coat using

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).



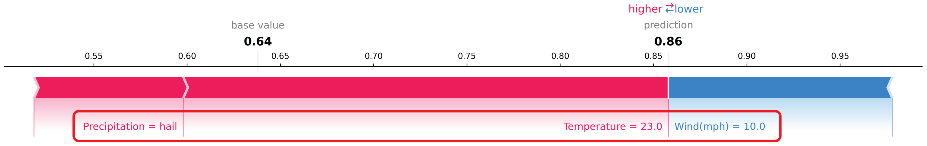
Let's take a closer look at this visual explanation.

Intro 4

Under the red and blue 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.

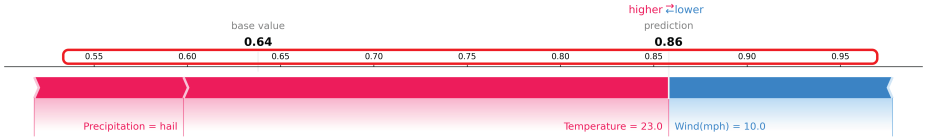
These factors can take inputs that are numerical (e.g., 30, 0) or categorical (e.g., rain, snow).



Intro 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).

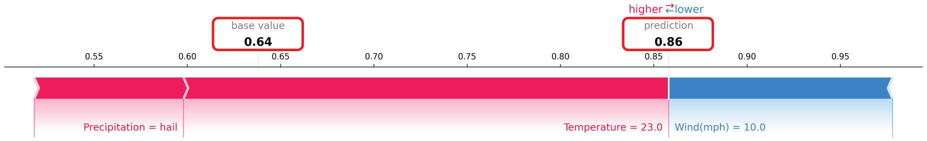


Intro 6

The **base value** represents the average value of the model's output across multiple predictions.

Imagine providing the model with a large set of different combinations of temperature, wind, and precipitation values, and asking the model to generate a prediction based on each combination. The model will generate probabilities such as 0.3, 0.4, 0.5, 0.6, 0.7, etc.

If we take the *average* of all the probabilities the model generates, we will get this **base value.**



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.

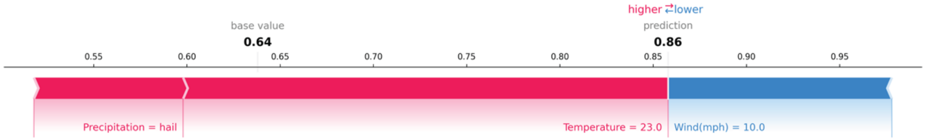
The visualization shows how, starting from the **base value**, each input values of temperature, wind, and precipitation can have a positive (red) contribution, pushing the prediction toward 'YES', or a negative (blue) contribution, pushing the prediction toward 'NO'.

Intro Test 1

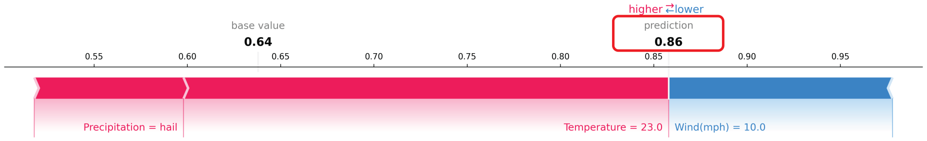
In the example below, what will the model predict?

○ YES, you should wear a coat

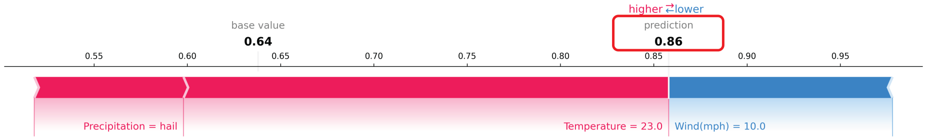
○ NO, do not wear a coat



Correct. In this case, the model prediction is 0.86, which is larger than 0.5, so the model will return YES.



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:

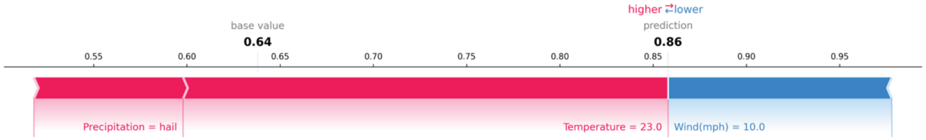
○ sleet

○ snow

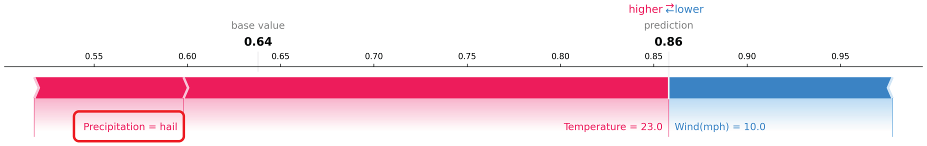
○ hail

○ rain

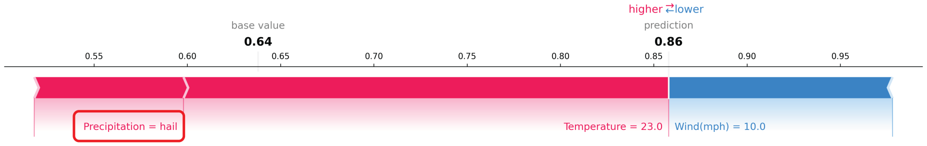
○ none



Correct - the value is printed next to the word **Precipitation** under the red and blue bars. This value is **hail**.



Not quite - the value is printed next to the word **Precipitation** under the red and blue bars. This value is **hail**.



By looking at the explanation image, please select the value for

wind speed input into the model:

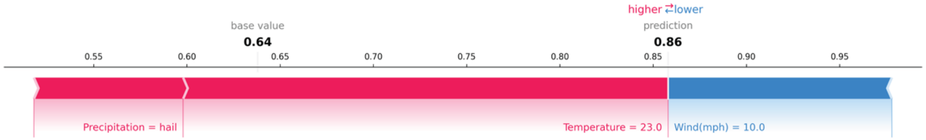
○ 20 mph

○ 0 mph

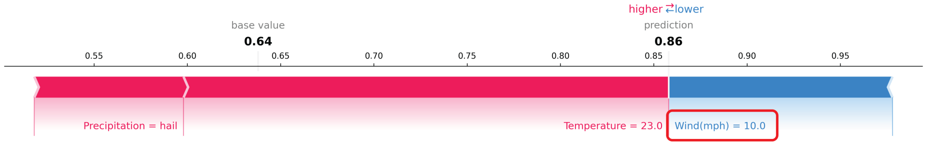
○ 10 mph

○ 5 mph

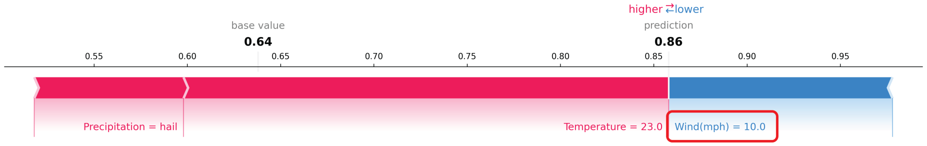
○ 15 mph



Correct - the value is printed next to the word **Wind(mph)** under the red and blue bars. This value is **10 mph**.

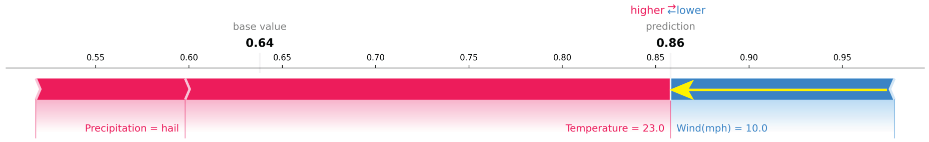


Not quite - the value is printed next to the word **Wind(mph)** under the red and blue bars. This value is **10 mph**.



Intro 8

Again, starting from the **base value** at the bottom, each input value of temperature, wind, and precipitation can push the model's **prediction** to be higher or lower.



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 **blue** and point to the *left*.

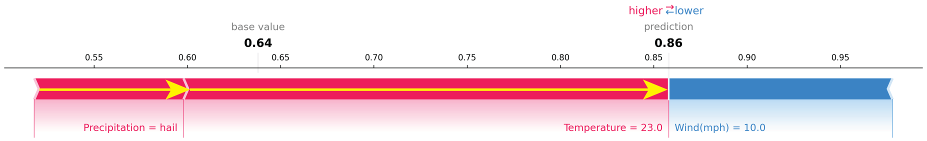
If the final prediction is pushed below 0.5, the model will return 'NO' (do not wear a coat).

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 **red** and always point to the *right*.

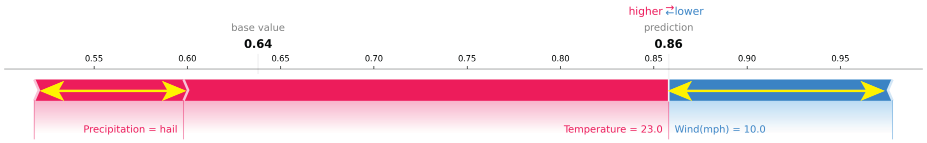
If the final prediction is pushed to 0.5 or above, the model will return 'YES' (wear a coat).



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.



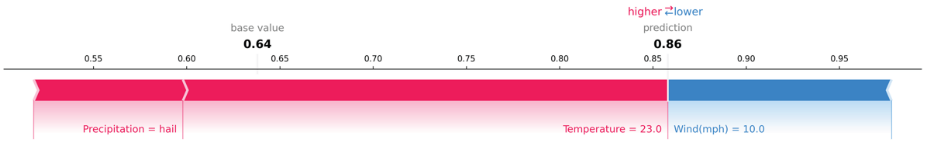
Intro Test 3

As a review, by looking at the explanation image, which factor(s) are pushing the model toward predicting 'YES'?

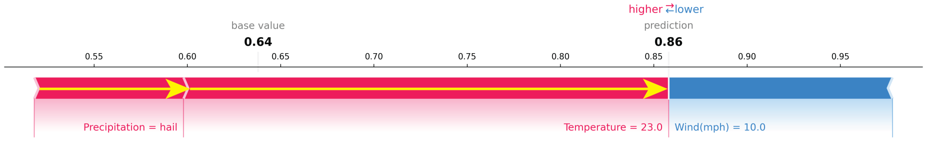
☐ Temperature

☐ Wind

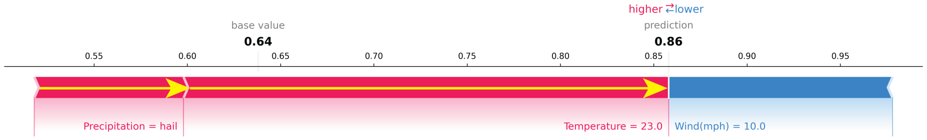
☐ Precipitation



Correct - In this case, the bars for temperature and precipitation are **red** and pointing to the **right**, so the values of these factors are pushing the prediction *higher* and pushing the model toward predicting 'YES'.



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

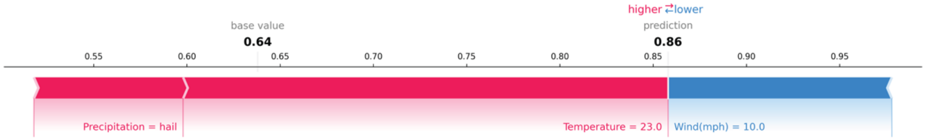


By looking at the explanation image, which factor(s) are pushing the model toward predicting 'NO'?

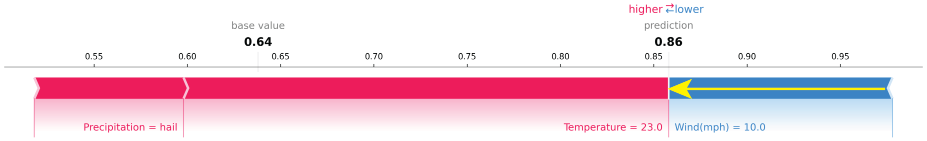
□ Temperature

□ Wind

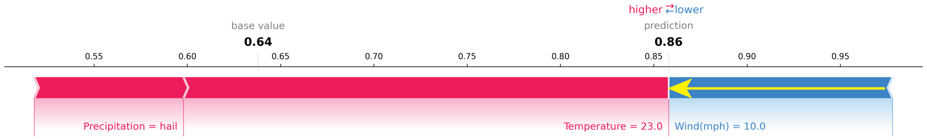
□ Precipitation



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

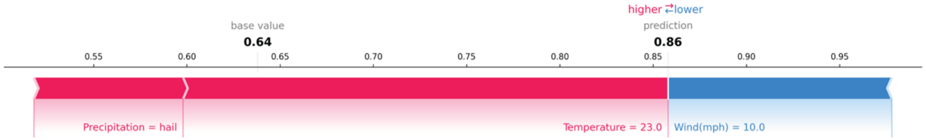


Which factor has the greatest predictive power?

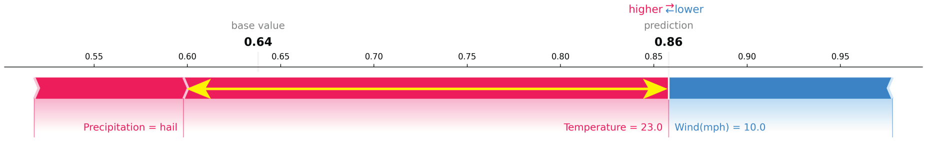
○ Temperature

○ Wind

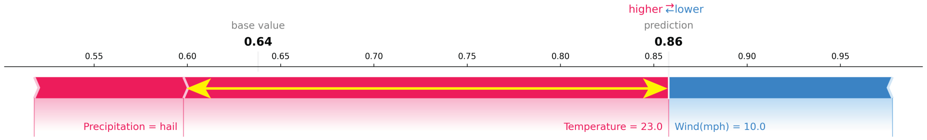
○ Precipitation



Correct - In this case, Temperature has the **longest** bar, so Temperature has the greatest predictive power.



Not quite - In this case, Temperature has the **longest** bar, so Temperature has the greatest predictive power.

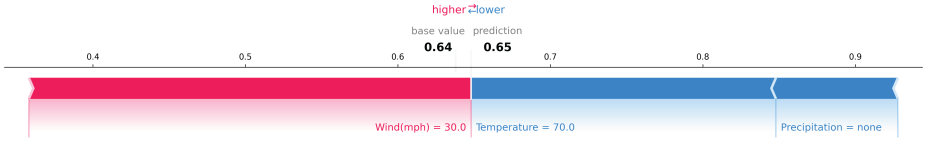


Intro Test 4

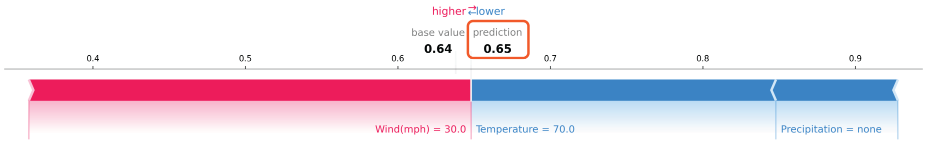
As a final review, what does the following model recommend you do?

○ YES, you should wear a coat

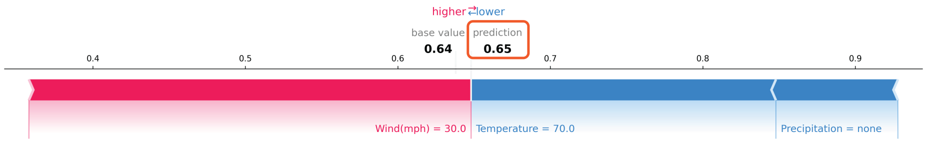
○ NO, do not wear a coat



Correct. In this case, the model prediction is 0.65, which is greater than 0.5, so the model will return 'YES'.



Incorrect. In this case, the model prediction is 0.65, which is greater than 0.5, so the model will return 'YES'.



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

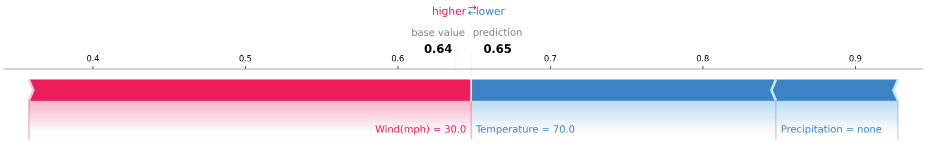
○ 84

○ 70

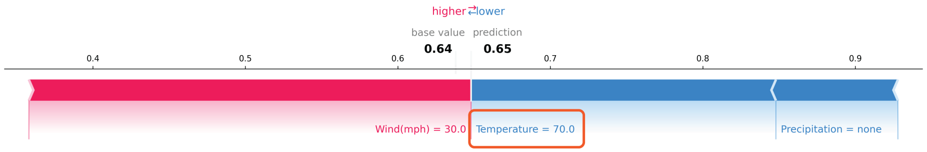
○ 61

○ 56

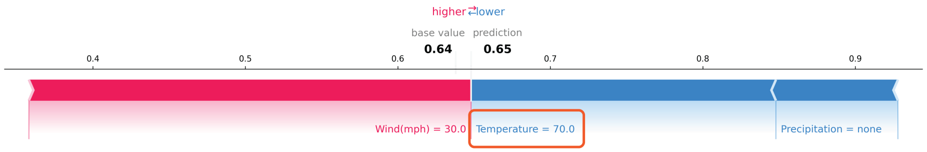
○ 37



Correct - the value is next to the word **Temperature** under the red and blue bars. This value is **70**.



Incorrect - the value is next to the word **Temperature** under the red and blue bars. This value is **70**.

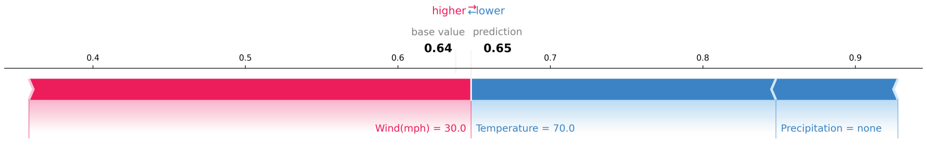


By looking at the explanation image, which factor(s) are pushing the model toward predicting 'NO?

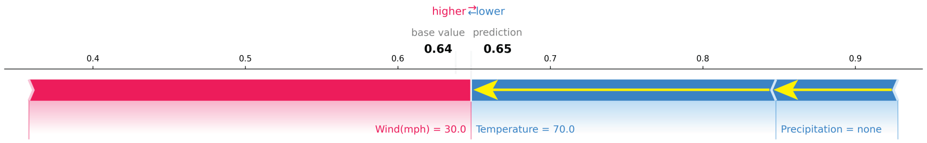
□ Temperature

□ Wind

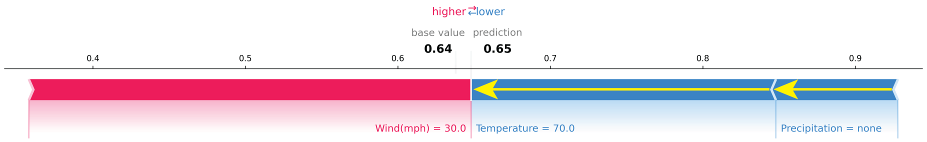
□ Precipitation



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

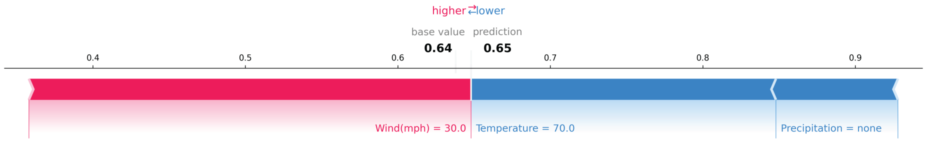


Which factor has the greatest predictive power?

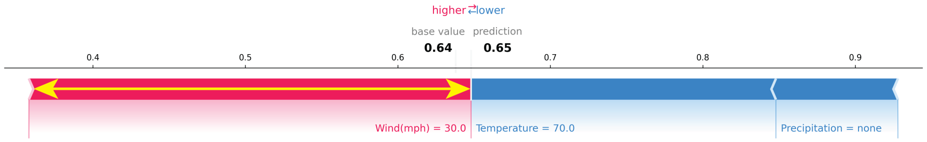
○ Temperature

○ Wind

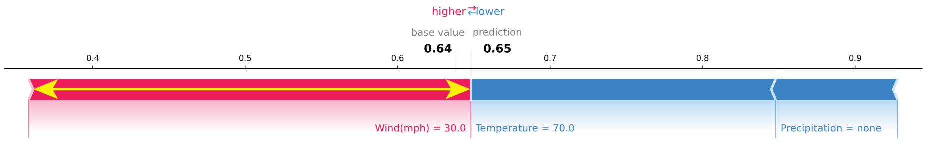
○ Precipitation



Correct - In this case, Wind(mph) has the **longest** bar, so Wind has the greatest predictive power.



Not quite - In this case, Wind(mph) has the **longest** bar, so Wind has the greatest predictive power.



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.

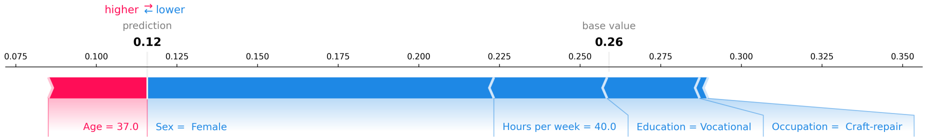
Woman 1

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).



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**?

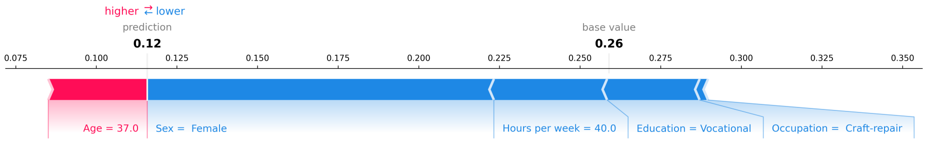
	Not at all	Very little	Somewhat	Moderately	A lot	A great deal
	1	2	3	4	5	6
Level of Trust						

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**?

	Not at all	Very little	Somewhat	Moderately	A lot	A great deal
	1	2	3	4	5	6
Level of Trust						

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.	○



When answering the previous questions about the given explanation, which design aspects of the visualization 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?

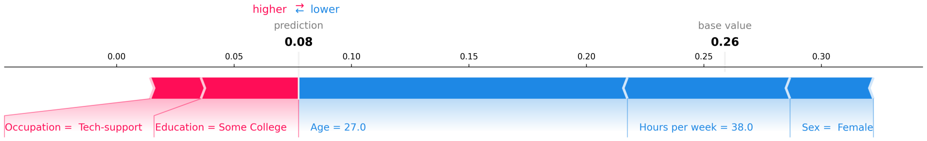
Woman 2

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).



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**?

	Not at all	Very little	Somewhat	Moderately	A lot	A great deal	
	1	2	3	4	5	6	
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**?

	Not at all	Very little	Somewhat	Moderately	A lot	A great deal
	1	2	3	4	5	6

Level of Trust

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.	○

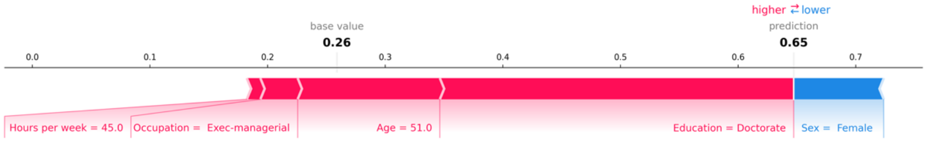
Woman 3

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

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**?

	Not at all	Very little	Somewhat	Moderately	A lot	A great deal
	1	2	3	4	5	6
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**?

	Not at all	Very little	Somewhat	Moderately	A lot	A great deal
	1	2	3	4	5	6
Level of Trust	○					□

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.	○

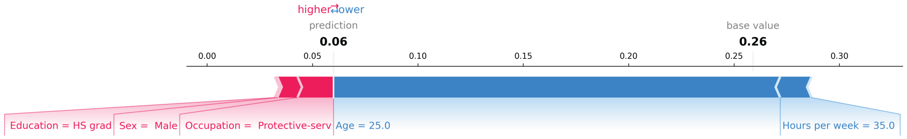
Man 1

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).



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

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
	1	2	3	4	5	6
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**?

	Not at all	Very little	Somewhat	Moderately	A lot	A great deal
	1	2	3	4	5	6
Level of Trust	○					□

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.	○

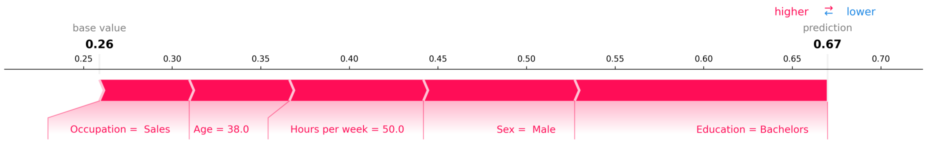
Man 2

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).



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

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	
	1	2	3	4	5	6	
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**?

	Not at all	Very little	Somewhat	Moderately	A lot	A great deal	
	1	2	3	4	5	6	
Level of Trust	○						□

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.	○

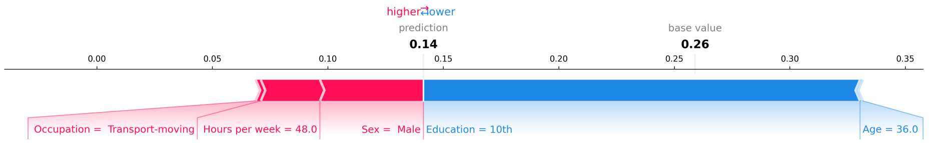
Man 3

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

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	
	1	2	3	4	5	6	
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**?

	Not at all	Very little	Somewhat	Moderately	A lot	A great deal	
	1	2	3	4	5	6	
Level of Trust							□

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?

Not well at all	Slightly well	Moderately well	Well	Very well	Extremely well
1	2	3	4	5	6

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

Not easy at all	Slightly easy	Moderately easy	Easy	Very easy	Extremely easy
1	2	3	4	5	6

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

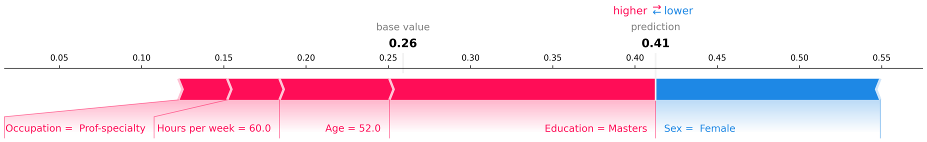
Not likely at all	Slightly likely	Moderately likely	Likely	Very likely	Extremely likely
1	2	3	4	5	6

Fairness

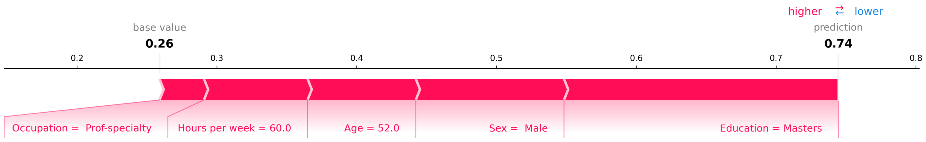
Below are two explanations for predictions made by the same loan approval machine learning model you have been seeing, for two people with almost identical features.

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).

Person A



Person B



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**?

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

	1	2	3	4	5	6
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**?

	Not at all 1	Very little 2	Somewhat 3	Moderately 4	A lot 5	A great deal 6
Level of Trust						□

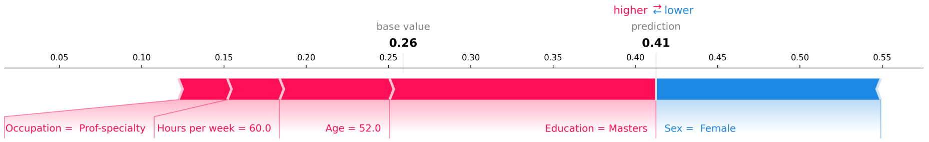
Please indicate whether you agree with the below statements.

	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.	○

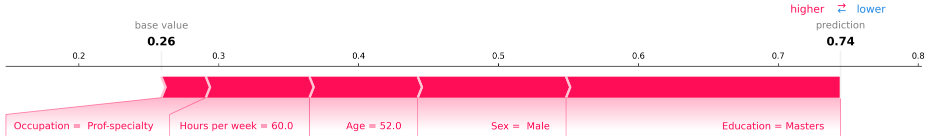
This model would probably give me a loan because I am similar to a person described in this question.	○
This model would probably give me a loan because I am different from a person described in this question.	○
This model would probably give me a loan because of previous decisions it has made.	○
This model probably would not give me a loan, and this would be the correct decision.	○

Fairness General

Person A



Person B



Do you think this model includes potentially discriminating factors?

○ YES

○ NO

If yes, which ones?

□ Age

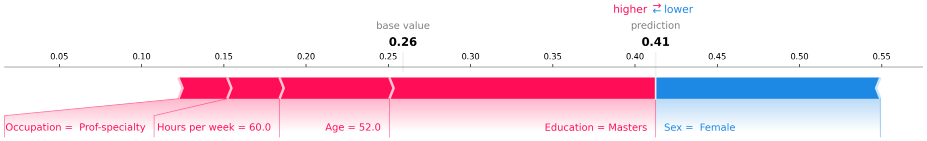
□ Hours Per Week

□ Education

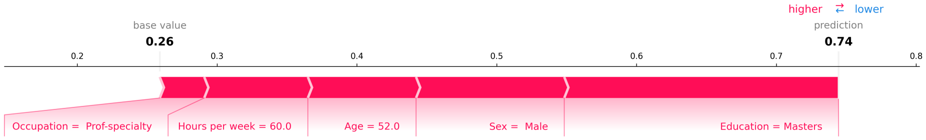
□ Occupation

□ Sex

Person A



Person B



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
○ Hawaiin/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