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).

y=YES (prediction **0.861**, score **1.547**) top features

Contribution	Feature	Value
+1.522	Temperature	23
+0.862	base value	1
+0.313	Precipitation	hail
-1.151	Wind(mph)	10

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

Intro 4

In the **Feature** and **Value** columns of the table, 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).

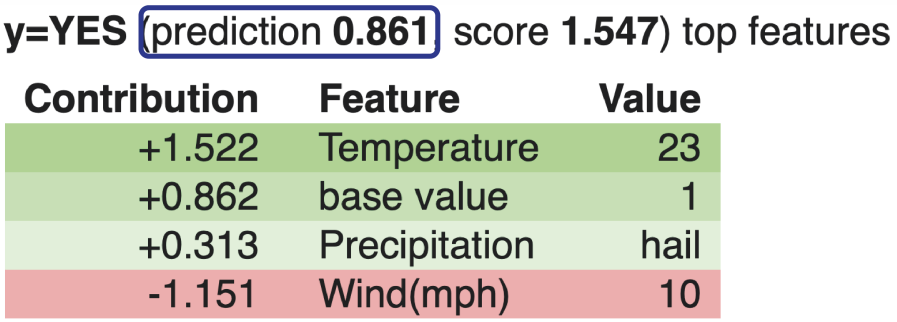
y=YES (prediction **0.861**, score **1.547**) top features

Contribution	Feature	Value
+1.522	Temperature	23
+0.862	base value	1
+0.313	Precipitation	hail
-1.151	Wind(mph)	10

Intro 5

The value next to the word "prediction" 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).

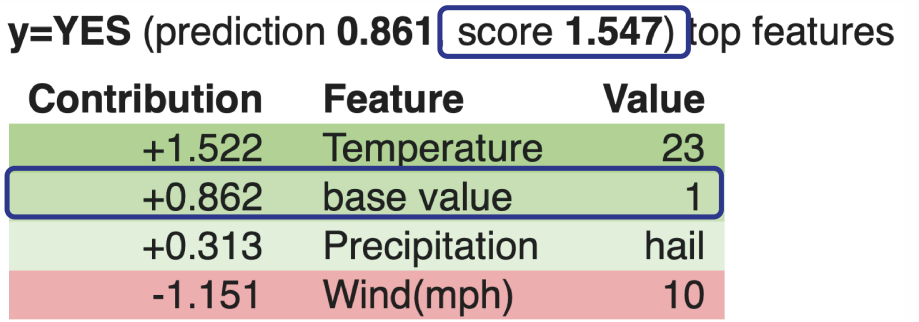


Intro 6

The explanation also has a **score** value and **base** value, calculated only for the purpose of explaining how each feature contributes to the final prediction. The **base value** represents the average value of the model's **score** 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 explanation algorithm will generate scores such as 0.3, 0.4, 0.5, 0.6, 0.7, etc.

If we take the *average* of all of these scores, we will get this **base value.**



The score is not the same as the prediction. You can put different values of temperature, wind, and precipitation into your model to generate a **prediction**.

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 above visualization shows each input values of temperature, wind, and precipitation can have a positive (green) contribution, pushing the prediction toward 'YES', or a negative (red) contribution, pushing the prediction toward 'NO'.

The score is the **sum** of these contributions and the base value.

Intro Test 1

In the example below, what will the model predict?

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

y=YES (prediction **0.861**, score **1.547**) top features

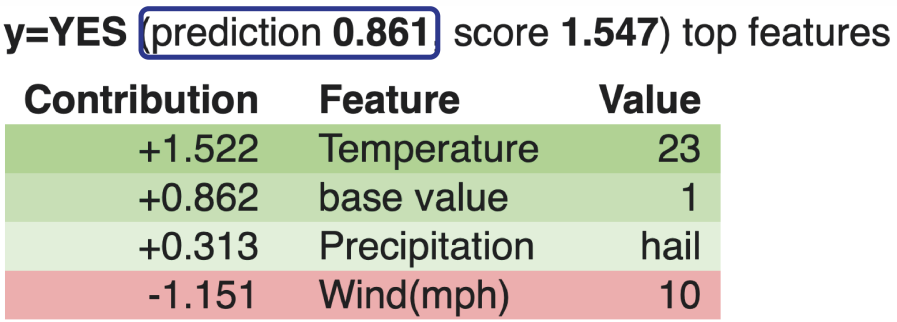
Contribution	Feature	Value
+1.522	Temperature	23
+0.862	base value	1
+0.313	Precipitation	hail
-1.151	Wind(mph)	10

Correct. In this case, the model prediction is 0.861, which is larger than 0.5, so the model will return YES.

y=YES (prediction **0.861**, score **1.547**) top features

Contribution	Feature	Value
+1.522	Temperature	23
+0.862	base value	1
+0.313	Precipitation	hail
-1.151	Wind(mph)	10

Not quite. In this case, the model prediction is 0.861, which is larger than 0.5, so the model will return YES.



y=YES (prediction **0.861**, score **1.547**) top features

Contribution	Feature	Value
+1.522	Temperature	23
+0.862	base value	1
+0.313	Precipitation	hail
-1.151	Wind(mph)	10

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

y=YES (prediction **0.861**, score **1.547**) top features

Contribution	Feature	Value
+1.522	Temperature	23
+0.862	base value	1
+0.313	Precipitation	hail
-1.151	Wind(mph)	10

Correct - the value is printed next to the **Precipitation** feature in the Value column. This value is **hail**.

y=YES (prediction **0.861**, score **1.547**) top features

Contribution	Feature	Value
+1.522	Temperature	23
+0.862	base value	1
+0.313	Precipitation	hail
-1.151	Wind(mph)	10

Not quite - the value is printed next to the **Precipitation** feature in the Value column. This value is **hail**.

y=YES (prediction **0.861**, score **1.547**) top features

Contribution	Feature	Value
+1.522	Temperature	23
+0.862	base value	1
+0.313	Precipitation	hail
-1.151	Wind(mph)	10

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

y=YES (prediction **0.861**, score **1.547**) top features

Contribution	Feature	Value
+1.522	Temperature	23
+0.862	base value	1
+0.313	Precipitation	hail
-1.151	Wind(mph)	10

Correct - the value is printed next to the **Wind(mph)** feature in the Value column. This value is **10mph**.

y=YES (prediction **0.861**, score **1.547**) top features

Contribution	Feature	Value
+1.522	Temperature	23
+0.862	base value	1
+0.313	Precipitation	hail
-1.151	Wind(mph)	10

Not Quite - the value is printed next to the **Wind(mph)** feature in the Value column. This value is **10mph**.

y=YES (prediction **0.861**, score **1.547**) top features

Contribution	Feature	Value
+1.522	Temperature	23
+0.862	base value	1
+0.313	Precipitation	hail
-1.151	Wind(mph)	10

Intro 8

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

y=YES (prediction **0.861**, score **1.547**) top features

Contribution	Feature	Value
+1.522	Temperature	23
+0.862	base value	1
+0.313	Precipitation	hail
-1.151	Wind(mph)	10

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 **red** and their contribution values are **negative**.

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 **green** and their contribution values are **positive**.

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

y=YES (prediction **0.861**, score **1.547**) top features

Contribution	Feature	Value
+1.522	Temperature	23
+0.862	base value	1
+0.313	Precipitation	hail
-1.151	Wind(mph)	10

Intro 10

The **absolute contribution value** (disregarding the sign) inside a row, its location, and the the depth of its color indicate the predictive power of a factor.

The rows are ordered from the most positive contribution to the most negative contribution. The features with the highest contributions are in the **top green row** or the **bottom red row**. Rows containing features with higher contributions are also a deeper green or red color.

The Wind(mph) feature is in the bottom row, but the Precipitation feature is not in the top row. Wind(mph) also has a higher absolute contribution value (1.151 vs 0.313). So Wind(mph) has a higher predictive power than Precipitation.

Never include the base value when comparing predictive power, since it is not a feature.

y=YES (prediction **0.861**, score **1.547**) top features

Contribution	Feature	Value
+1.522	Temperature	23
+0.862	base value	1
+0.313	Precipitation	hail
-1.151	Wind(mph)	10

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

y=YES (prediction **0.861**, score **1.547**) top features

Contribution	Feature	Value
+1.522	Temperature	23
+0.862	base value	1
+0.313	Precipitation	hail
-1.151	Wind(mph)	10

Correct - In this case, the rows for temperature and precipitation are **green** and their contribution values are **positive**, so the values of these factors are pushing the prediction *higher* and pushing the model toward predicting 'YES'.

y=YES (prediction **0.861**, score **1.547**) top features

Contribution	Feature	Value
+1.522	Temperature	23
+0.862	base value	1
+0.313	Precipitation	hail
-1.151	Wind(mph)	10

Not quite - In this case, the bars for temperature and precipitation are **green** and their contribution values are **positive,** so the values of these factors are pushing the prediction *higher* and pushing the model toward predicting 'YES'.

y=YES (prediction **0.861**, score **1.547**) top features

Contribution	Feature	Value
+1.522	Temperature	23
+0.862	base value	1
+0.313	Precipitation	hail
-1.151	Wind(mph)	10

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

- [ ] Temperature
- [ ] Wind
- [ ] Precipitation

y=YES (prediction **0.861**, score **1.547**) top features

Contribution	Feature	Value
+1.522	Temperature	23
+0.862	base value	1
+0.313	Precipitation	hail
-1.151	Wind(mph)	10

Correct - In this case, the row for Wind(mph) is **red** and its contribution value is **negative**, so the value of this factor is pushing the prediction *lower* and pushing the model toward predicting 'NO'.

y=YES (prediction **0.861**, score **1.547**) top features

Contribution	Feature	Value
+1.522	Temperature	23
+0.862	base value	1
+0.313	Precipitation	hail
-1.151	Wind(mph)	10

Not quite - In this case, the row for Wind(mph) is red and its contribution value is **negative** , so the value of this factor is pushing the prediction *lower* and pushing the model toward predicting 'NO'.

y=YES (prediction **0.861**, score **1.547**) top features

Contribution	Feature	Value
+1.522	Temperature	23
+0.862	base value	1
+0.313	Precipitation	hail
-1.151	Wind(mph)	10

Which factor has the greatest predictive power?

- ○ Temperature
- ○ Wind
- ○ Precipitation

y=YES (prediction **0.861**, score **1.547**) top features

Contribution	Feature	Value
+1.522	Temperature	23
+0.862	base value	1
+0.313	Precipitation	hail
-1.151	Wind(mph)	10

Correct - In this case, Temperature has the highest absolute contribution value, it is the top bar, and it has the deepest color. So Temperature has the greatest predictive power.

y=YES (prediction **0.861**, score **1.547**) top features

Contribution	Feature	Value
+1.522	Temperature	23
+0.862	base value	1
+0.313	Precipitation	hail
-1.151	Wind(mph)	10

Not quite - In this case, Temperature has the highest absolute contribution value, it is the top bar, and it has the deepest color. So Temperature has the greatest predictive power.

y=YES (prediction **0.861**, score **1.547**) top features

Contribution	Feature	Value
+1.522	Temperature	23
+0.862	base value	1
+0.313	Precipitation	hail
-1.151	Wind(mph)	10

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

y=YES (prediction **0.648**, score **0.589**) top features

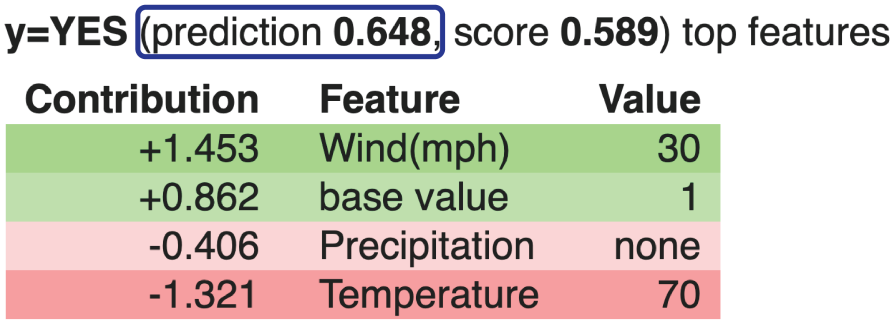
Contribution	Feature	Value
+1.453	Wind(mph)	30
+0.862	base value	1
-0.406	Precipitation	none
-1.321	Temperature	70

Correct. In this case, the model prediction is 0.648, which is greater than 0.5, so the model will return 'YES'.

y=YES (prediction **0.648**, score **0.589**) top features

Contribution	Feature	Value
+1.453	Wind(mph)	30
+0.862	base value	1
-0.406	Precipitation	none
-1.321	Temperature	70

Incorrect. In this case, the model prediction is 0.648, which is greater than 0.5, so the model will return 'YES'.



y=YES (prediction **0.648**, score **0.589**) top features

Contribution	Feature	Value
+1.453	Wind(mph)	30
+0.862	base value	1
-0.406	Precipitation	none
-1.321	Temperature	70

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

- ○ 84
- ○ 70
- ○ 61
- ○ 56

○ 37

y=YES (prediction **0.648**, score **0.589**) top features

Contribution	Feature	Value
+1.453	Wind(mph)	30
+0.862	base value	1
-0.406	Precipitation	none
-1.321	Temperature	70

Correct - the value is in the **Value** column next to the **Temperature** feature. This value is **70**.

y=YES (prediction **0.648**, score **0.589**) top features

Contribution	Feature	Value
+1.453	Wind(mph)	30
+0.862	base value	1
-0.406	Precipitation	none
-1.321	Temperature	70

Incorrect - the value is in the **Value** column next to the **Temperature** feature. This value is **70**.

y=YES (prediction **0.648**, score **0.589**) top features

Contribution	Feature	Value
+1.453	Wind(mph)	30
+0.862	base value	1
-0.406	Precipitation	none
-1.321	Temperature	70

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

- ☐ Temperature
- ☐ Wind
- ☐ Precipitation

y=YES (prediction **0.648**, score **0.589**) top features

Contribution	Feature	Value
+1.453	Wind(mph)	30
+0.862	base value	1
-0.406	Precipitation	none
-1.321	Temperature	70

Correct - In this case, the bars for Temperature and Precipitation **red** and their contribution value is **negative**, so the value of these factors is pushing the prediction *lower* and pushing the

model toward predicting 'NO'.

y=YES (prediction **0.648**, score **0.589**) top features

Contribution	Feature	Value
+1.453	Wind(mph)	30
+0.862	base value	1
-0.406	Precipitation	none
-1.321	Temperature	70

Not quite - In this case, the bars for Temperature and Precipitation **red** and their contribution value is **negative**, so the value of these factors is pushing the prediction *lower* and pushing the model toward predicting 'NO'.

y=YES (prediction **0.648**, score **0.589**) top features

Contribution	Feature	Value
+1.453	Wind(mph)	30
+0.862	base value	1
-0.406	Precipitation	none
-1.321	Temperature	70

Which factor has the greatest predictive power?

- ○ Temperature
- ○ Wind
- ○ Precipitation

y=YES (prediction **0.648**, score **0.589**) top features

Contribution	Feature	Value
+1.453	Wind(mph)	30
+0.862	base value	1
-0.406	Precipitation	none
-1.321	Temperature	70

Correct - In this case, Wind(mph) has the highest absolute contribution value, it is the top bar, and it has the deepest color. So Wind has the greatest predictive power.

y=YES (prediction **0.648**, score **0.589**) top features

Contribution	Feature	Value
+1.453	Wind(mph)	30
+0.862	base value	1
-0.406	Precipitation	none
-1.321	Temperature	70

Not quite - In this case, Wind(mph) has the highest contribution value, it is the top bar, and it has the deepest color. So Wind has the greatest predictive power.

y=YES (prediction **0.648**, score **0.589**) top features

Contribution	Feature	Value
+1.453	Wind(mph)	30
+0.862	base value	1
-0.406	Precipitation	none
-1.321	Temperature	70

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).

y=YES (prediction **0.116**, score **-2.032**) top features

Contribution	Feature	Value
+2.284	Age	37.0
-0.001	base value	1
-0.193	Occupation	Craft-repair
-2.135	Education	Vocational
-2.611	Hours per week	40.0
-7.996	Sex	Female

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.	○

y=YES (prediction **0.116**, score **-2.032**) top features

Contribution	Feature	Value
+2.284	Age	37.0
-0.001	base value	1
-0.193	Occupation	Craft-repair
-2.135	Education	Vocational
-2.611	Hours per week	40.0
-7.996	Sex	Female

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).

y=YES (prediction **0.077**, score **-2.477**) top features

Contribution	Feature	Value
+0.990	Education	Some College
+0.514	Occupation	Tech-support
-0.001	base value	1
-0.845	Sex	Female
-1.647	Hours per week	38.0
-3.333	Age	27.0

\\

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).

y=YES (prediction **0.647**, score **0.607**) top features

Contribution	Feature	Value
+0.177	Education	Doctorate
+0.071	Age	51
+0.019	Occupation	Exec-managerial
+0.007	Hours per week	45.0
-0.001	base value	1
-0.045	Sex	Female

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**?

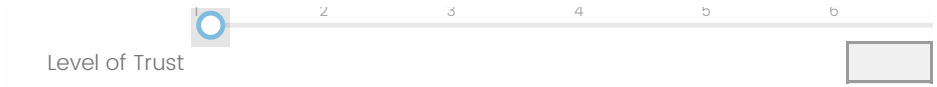
Not at all | Very little | Somewhat | Moderately | A lot | A great deal

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



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).

y=YES (prediction **0.060**, score **-2.749**) top features

Contribution	Feature	Value
+0.338	Occupation	Protective-serv
+0.190	Sex	Male
+0.022	Education	HS grad
-0.001	base value	1
-0.295	Hours per week	35.0
-4.192	Age	25.0

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

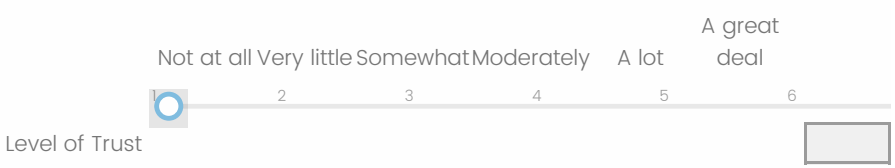
Which factor(s) are pushing the model toward predicting 'NO'?

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

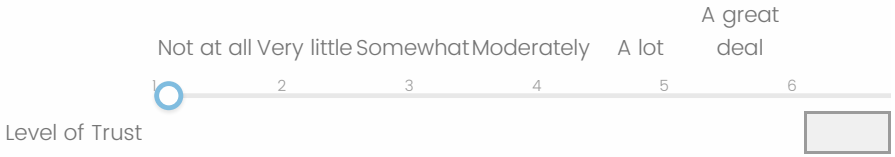
Which factor(s) are pushing the model toward predicting 'YES'?

- ☐ 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**?



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.	○

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).

y=YES (prediction **0.670**, score **0.708**) top features

Contribution	Feature	Value
+0.094	Education	Bachelors
+0.056	Sex	Male
+0.050	Hours per week	50.0
+0.037	Age	38.0
+0.033	Occupation	Sales
-0.001	base value	1

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

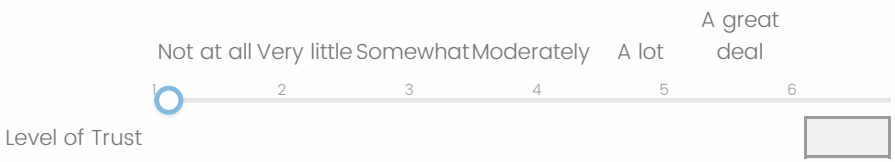
Which factor(s) are pushing the model toward predicting 'NO'?

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

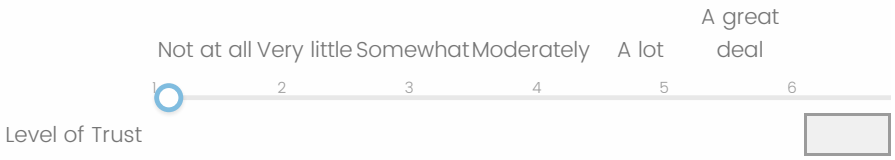
Which factor(s) are pushing the model toward predicting 'YES'?

- ☐ 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**?



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.	○

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).

y=YES (prediction **0.141**, score **-1.803**) top features

Contribution	Feature	Value
+3.393	Sex	Male
+1.939	Hours per week	48.0
+0.167	Occupation	Transport-moving
-0.001	base value	1
-0.091	Age	36.0
-14.313	Education	10th

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

y=YES (prediction **0.413**, score **-0.353**) top features

Contribution	Feature	Value
+0.101	Education	Masters
+0.042	Age	52.0
+0.020	Hours per week	60.0
+0.018	Occupation	Prof-specialty
-0.001	base value	1
-0.085	Sex	Female

Person B

y=YES (prediction **0.745**, score **170**) top features

Contribution	Feature	Value
+0.171	Education	Masters
+0.092	Sex	Male
+0.067	Age	52.0
+0.065	Hours per week	60.0
+0.028	Occupation	Prof-specialty
-0.001	base value	1

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	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 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

y=YES (prediction **0.413**, score **-0.353**) top features

Contribution	Feature	Value
+0.101	Education	Masters
+0.042	Age	52.0
+0.020	Hours per week	60.0
+0.018	Occupation	Prof-specialty
-0.001	base value	1
-0.085	Sex	Female

Person B

y=YES (prediction **0.745**, score **170**) top features

Contribution	Feature	Value
+0.171	Education	Masters
+0.092	Sex	Male
+0.067	Age	52.0
+0.065	Hours per week	60.0
+0.028	Occupation	Prof-specialty
-0.001	base value	1

Do you think this model includes potentially discriminating factors?

- ○ YES
- ○ NO

If yes, which ones?

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

Person A

y=YES (prediction **0.413**, score **-0.353**) top features

Contribution	Feature	Value
+0.101	Education	Masters
+0.042	Age	52.0
+0.020	Hours per week	60.0
+0.018	Occupation	Prof-specialty
-0.001	base value	1
-0.085	Sex	Female

Person B

y=YES (prediction **0.745**, score **170**) top features

Contribution	Feature	Value
+0.171	Education	Masters
+0.092	Sex	Male
+0.067	Age	52.0
+0.065	Hours per week	60.0
+0.028	Occupation	Prof-specialty
-0.001	base value	1

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