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

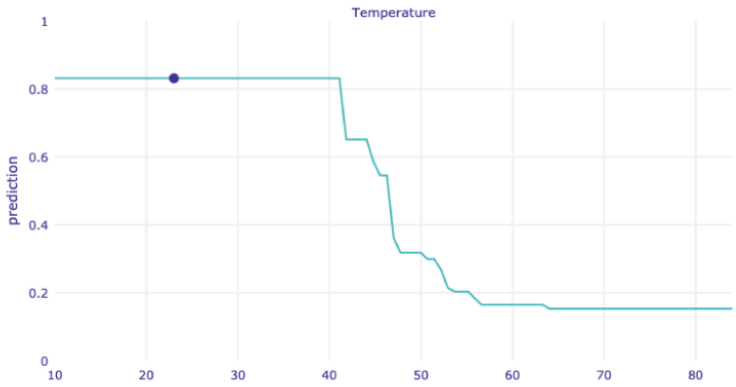
the current temperature, wind speed, and precipitation.

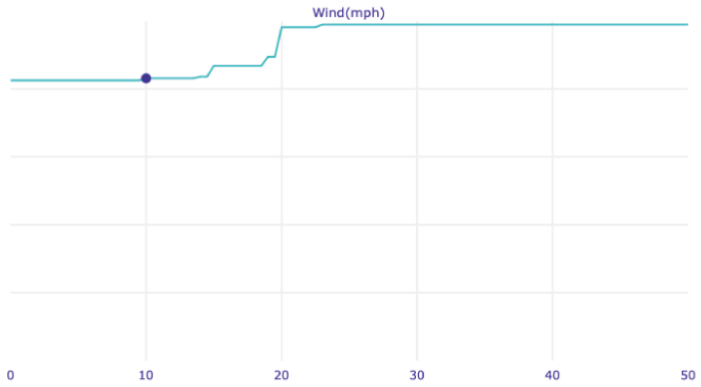
If that probability is greater than or equal to 0.5, then the model will recommend that you put on a coat. If the probability is less than 0.5, then the model will recommend that you do NOT put on a coat.

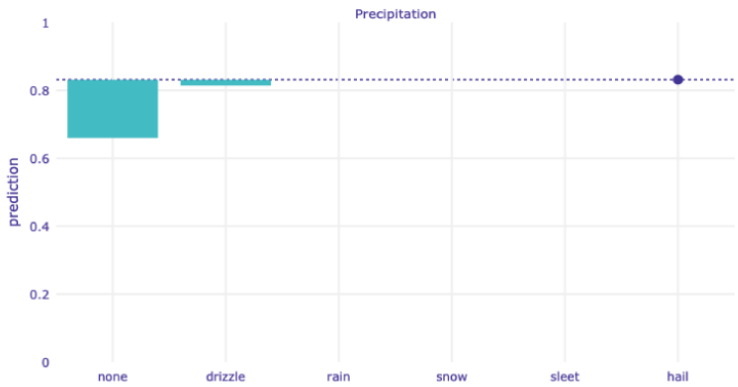
Intro 3

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).

Each chart in this explanation shows you how the model's prediction will change if **only** that feature is changed.







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

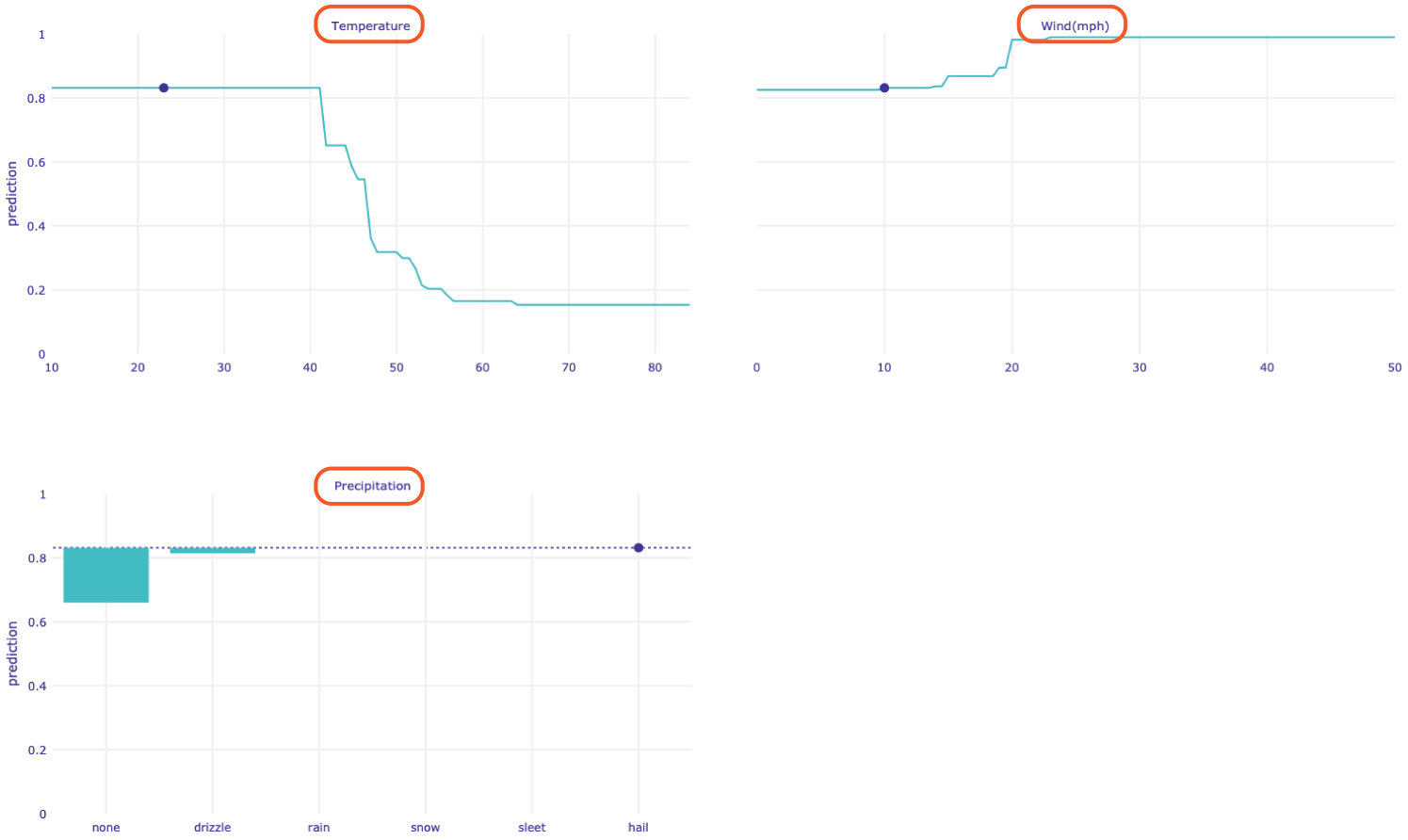
Intro 4

You can see two types of charts below. The first two are line charts, and the second is a bar chart.

Each chart has the name of the feature it corresponds to listed above.

Line charts correspond to **numerical** features - features that can take on a range of numerical values.

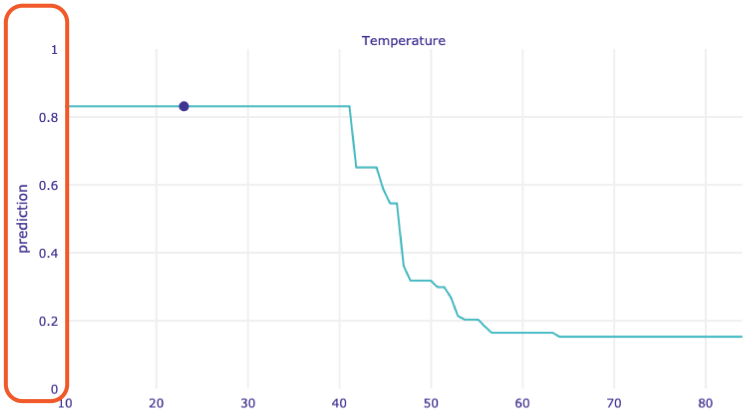
Bar charts correspond to **categorical** features - features which have values that are not numbers.

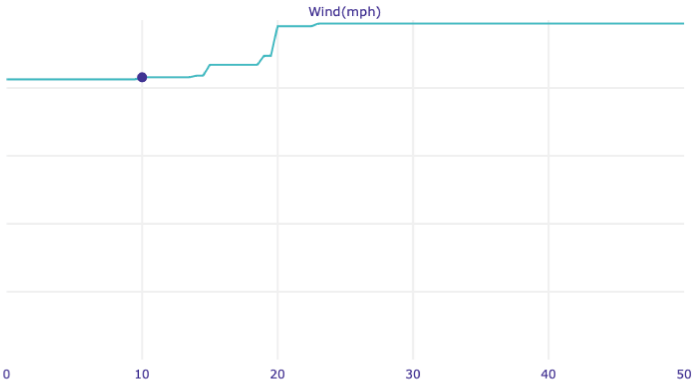


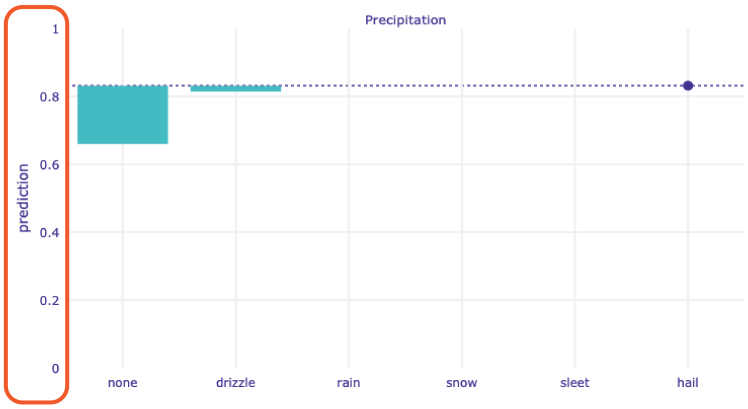
Intro 5

The vertical axis on both the line and the bar charts shows the probability value predicted 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).



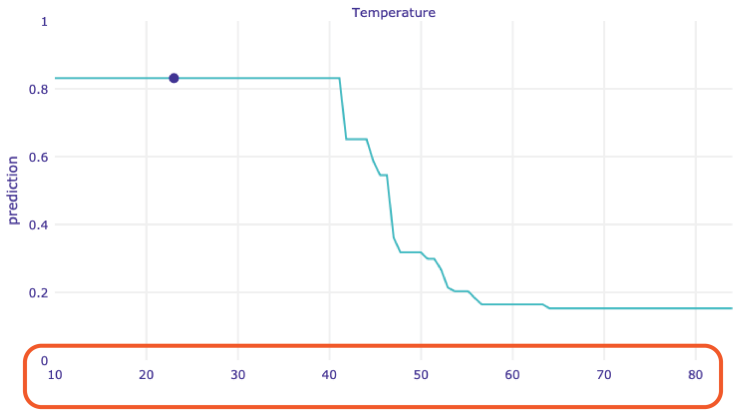


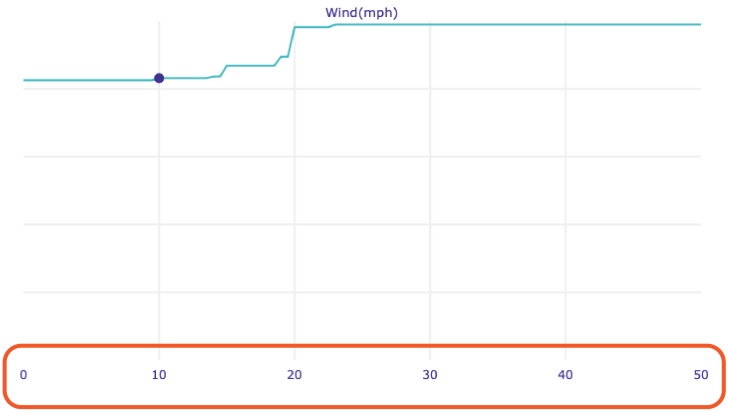


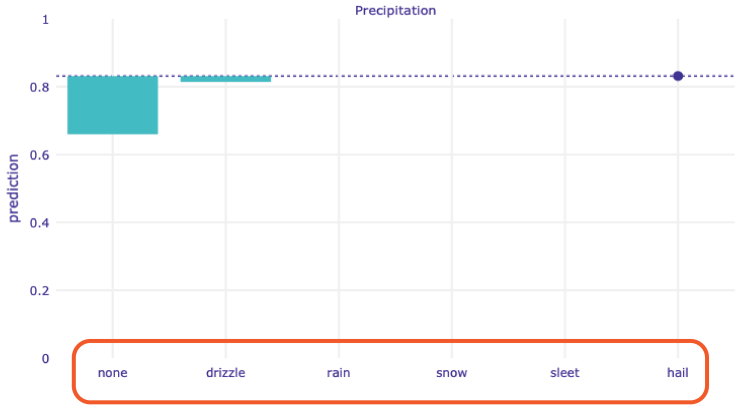
Intro 6

The horizontal axis on both the line and the bar charts shows the possible values for each feature. For example, the x axis for

temperature ranges from 10 to 80, because that is the range of values that temperature can take for this model. Since precipitation is categorical, it has six possible values - none, drizzle, rain, snow, sleet, and hail.







Intro 7

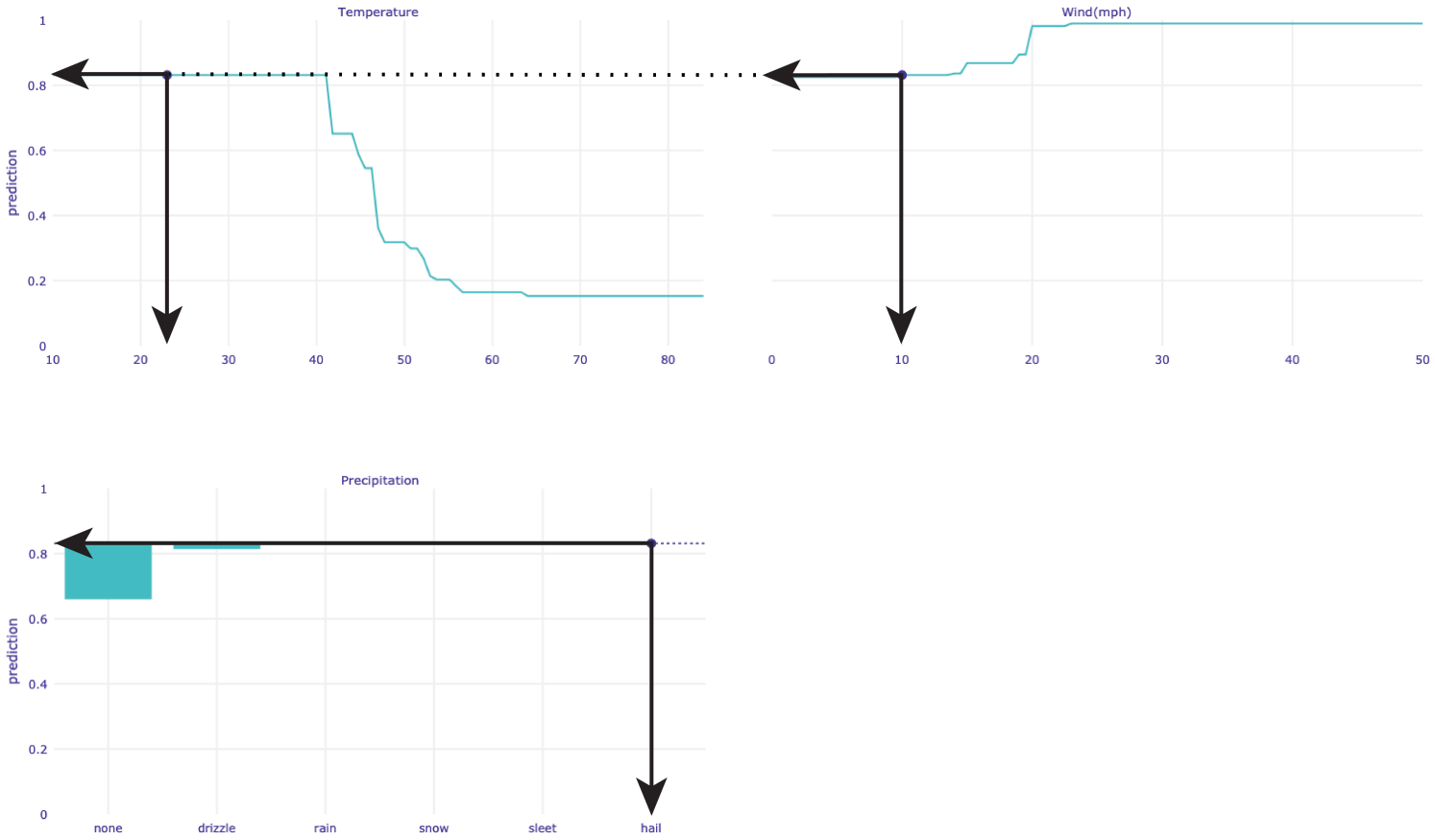
You can put different values of temperature, wind, and precipitation into your model to generate a model prediction.

The position of blue dot in each chart shows the the value of each feature which was put in to generate the **current prediction** (x-axis), and value of the prediction itself (y-axis).

The horizontal position of the dot is the value of the feature. You can look **below the dot** to see what the value of temperature, wind or precipitation is. The vertical position is the resulting model prediction - you can look **to the left of the dot** to see what the model predicted.

Here, our feature values are:
Temperature = 23
Wind(mph) = 10
Precipitation = hail

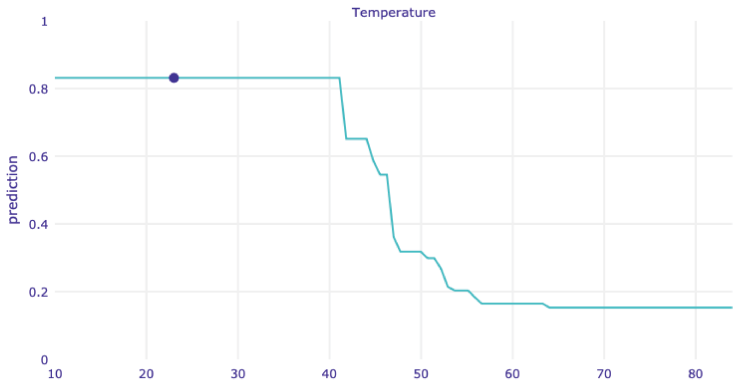
You are seeing a description for a **single model prediction**, so the y-axis prediction value where the blue dot is will be **the same** for every feature. Here, that value is 0.847.

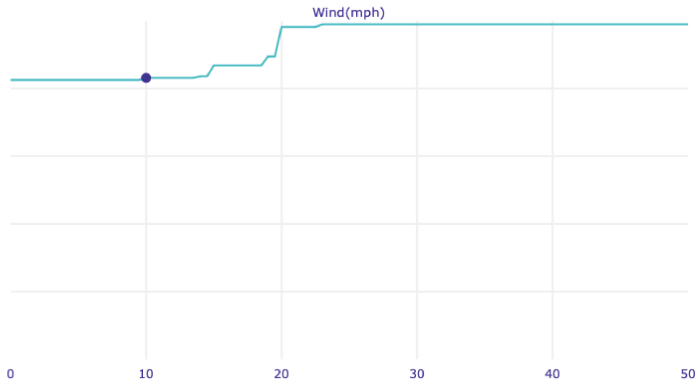


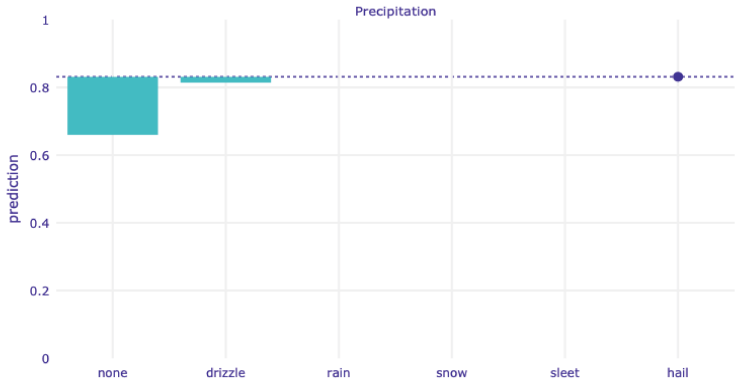
Intro Test 1

In the example below, what will the model predict?

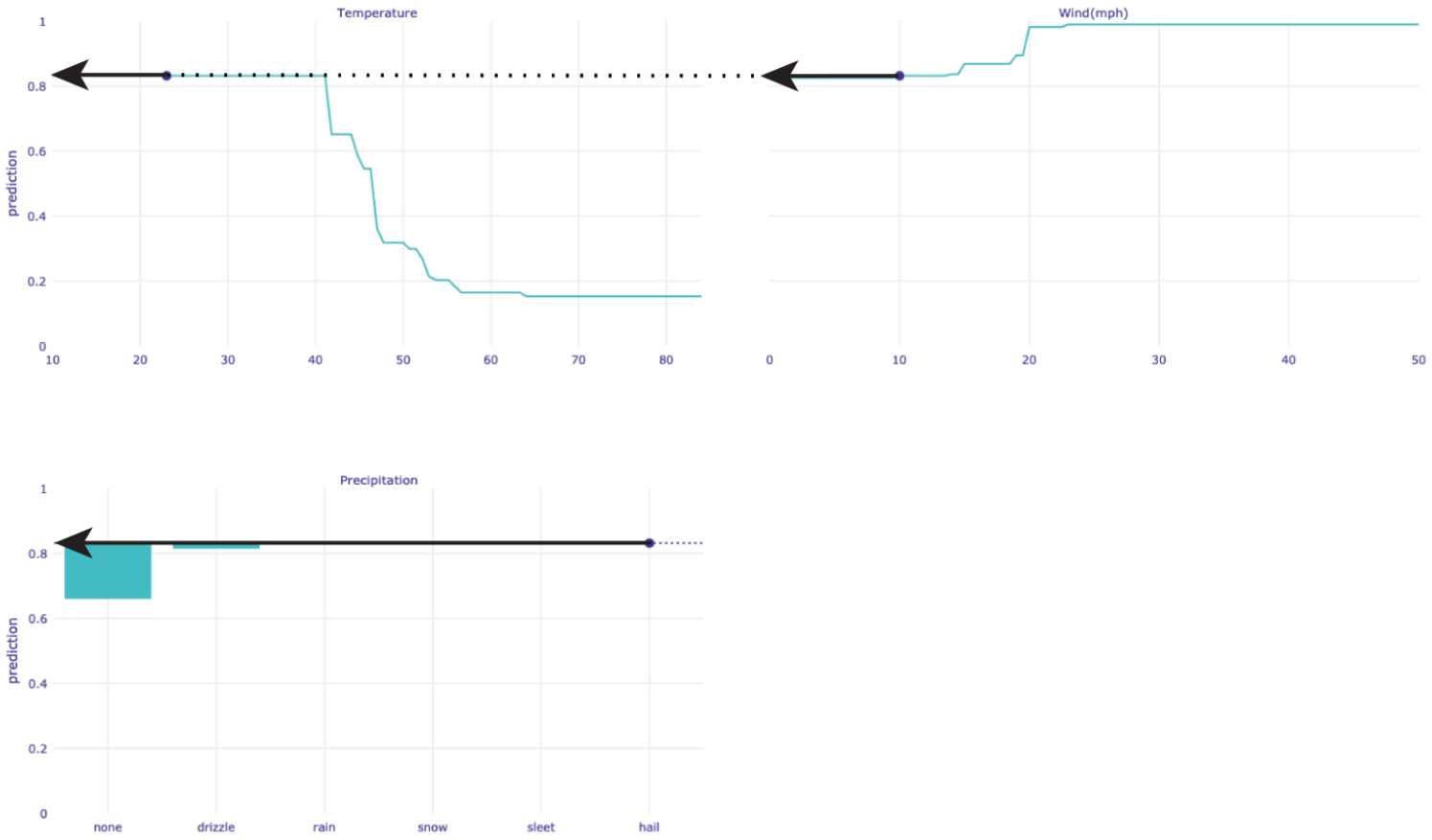
- ○ YES, you should wear a coat
- ○ NO, do not wear a coat



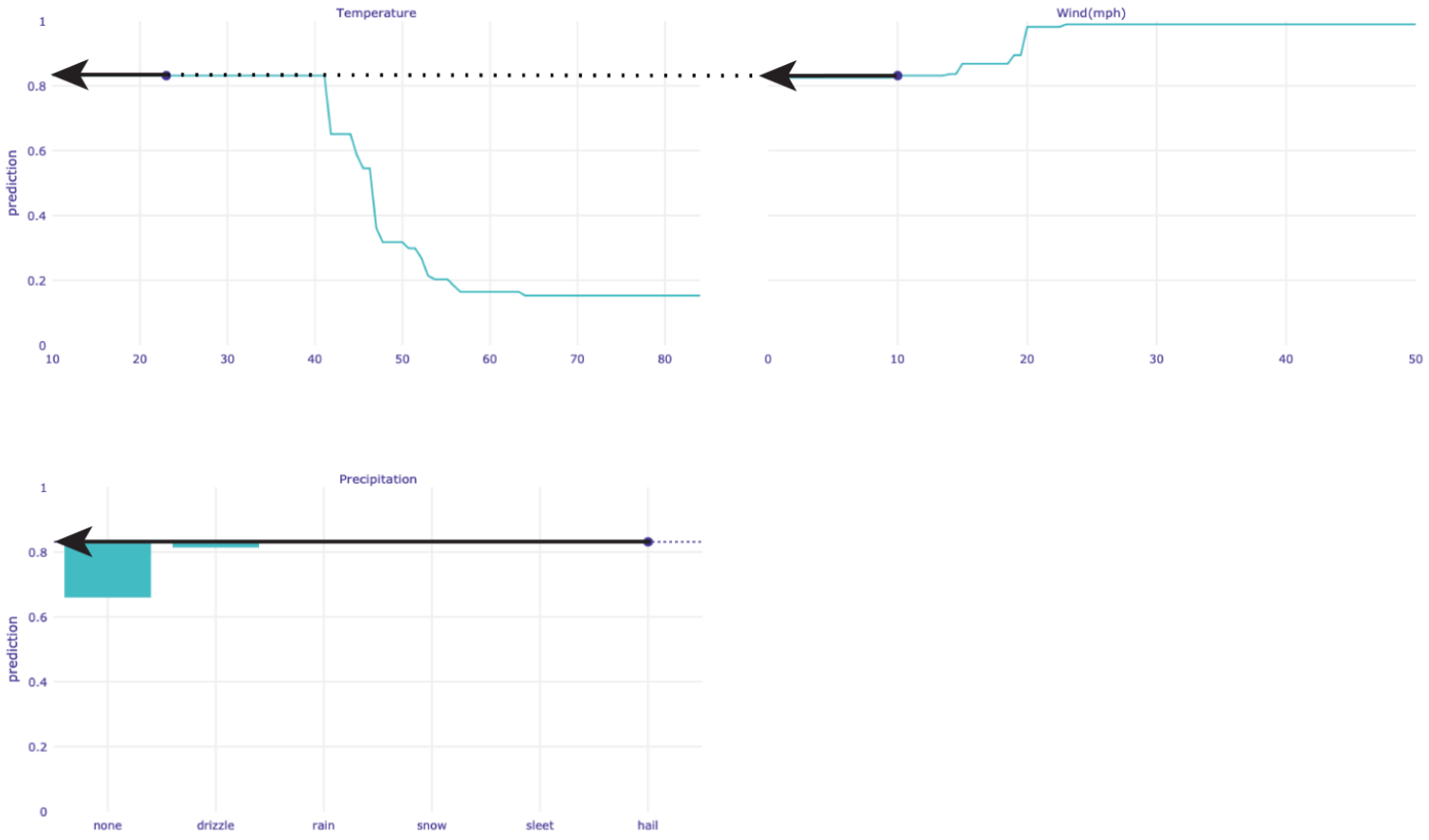




Correct. In this case, the model prediction is 0.847, which is larger than 0.5, so the model will return YES.



Not quite. In this case, the model prediction is 0.847, which is larger than 0.5, so the model will return YES.

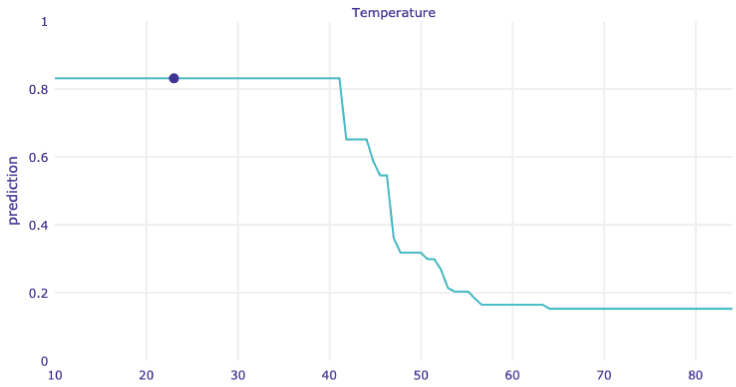


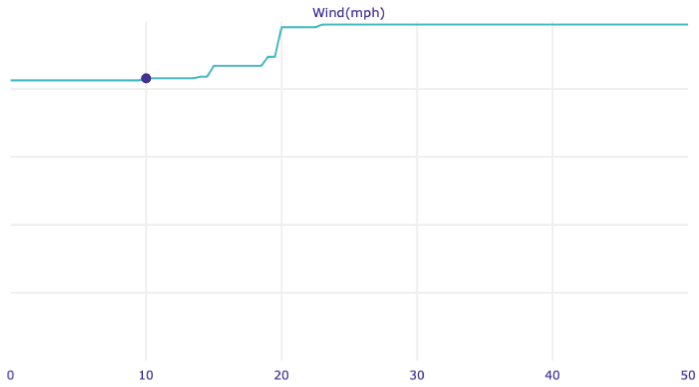
Intro Test 2

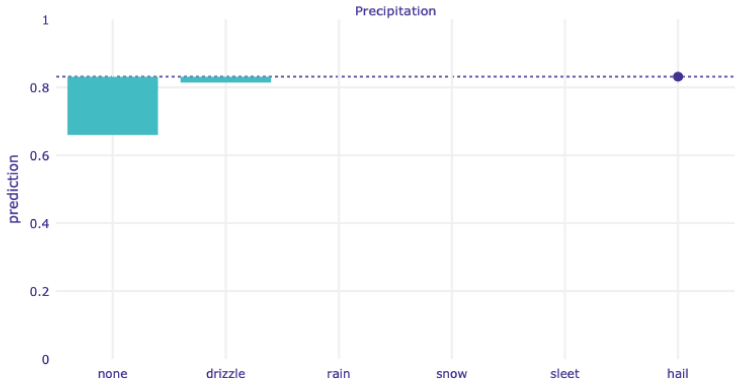
As another review, by looking at the explanation image, please select the value for **precipitation** input into the model:

○ sleet

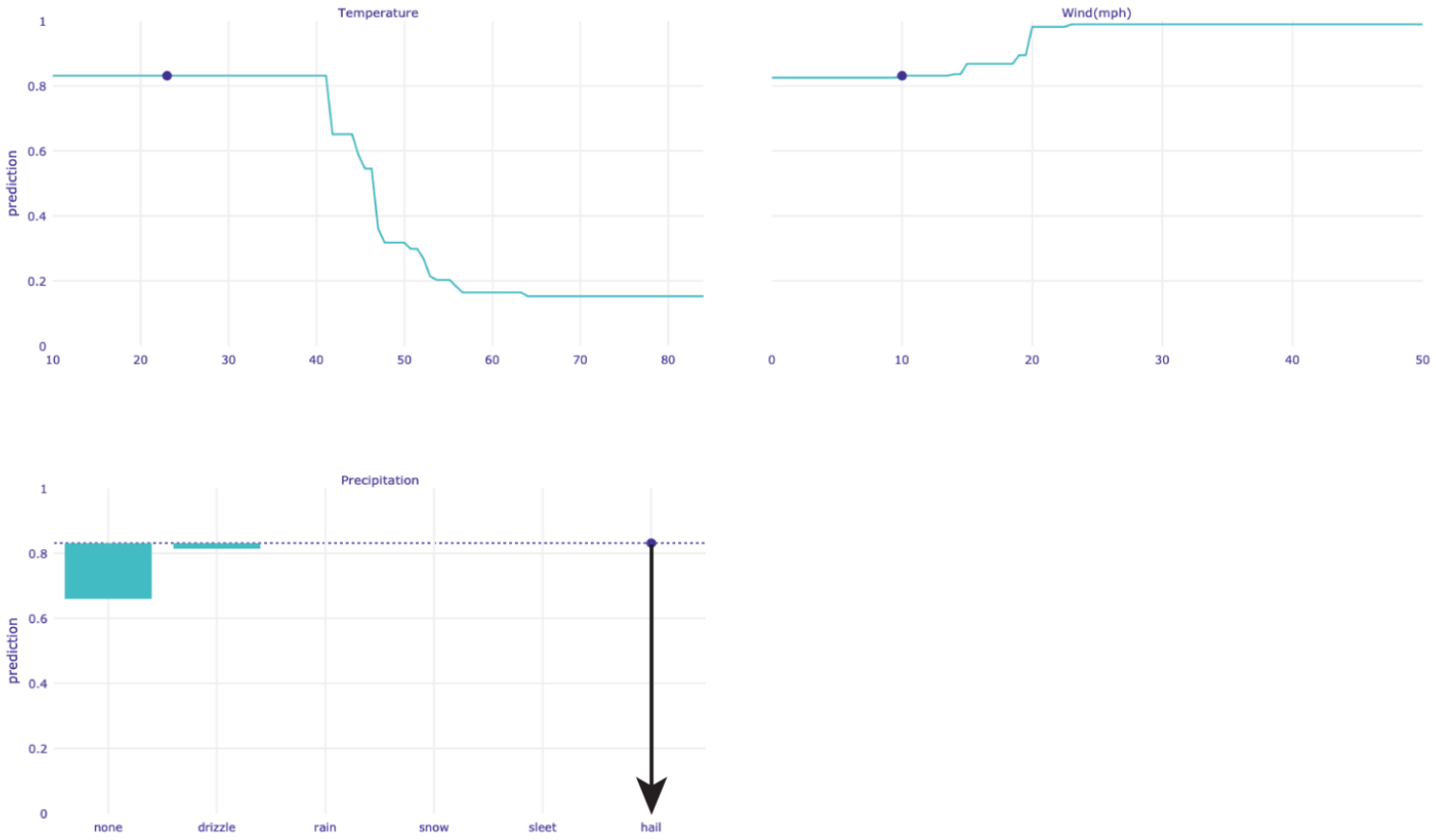
- snow
- hail
- rain
- none



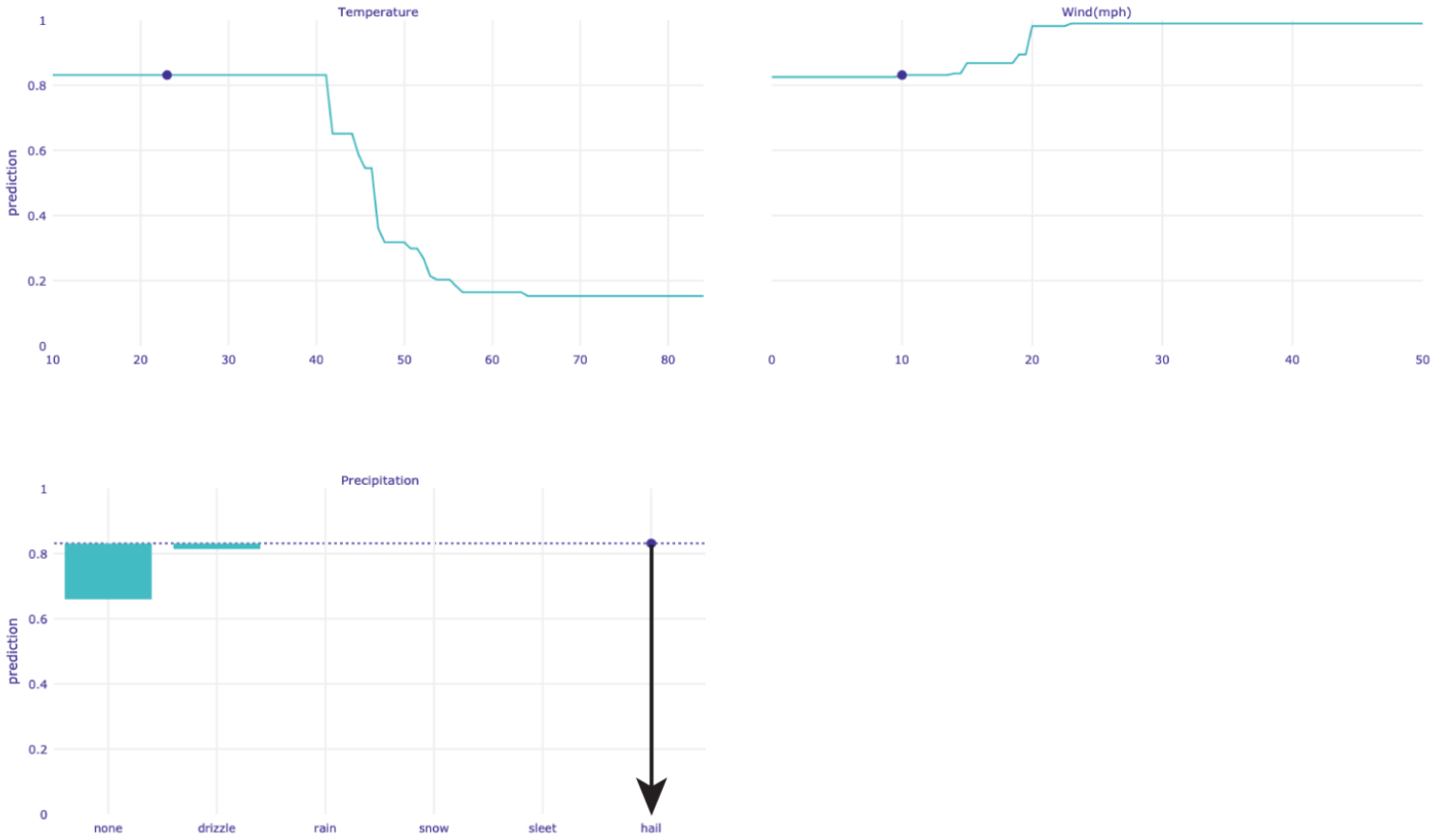




Correct - the value is below the blue dot in the precipitation graph. This value is **hail**.



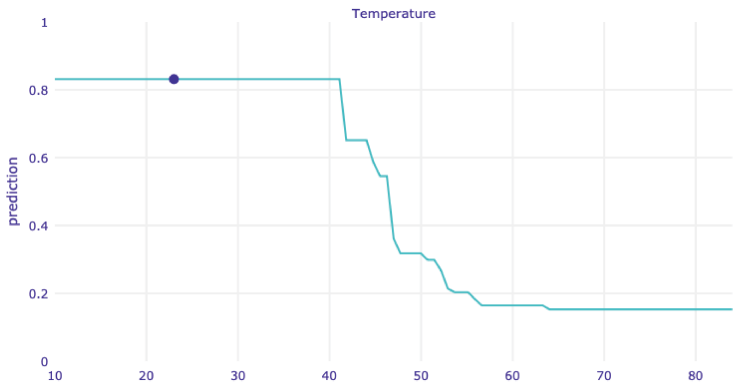
Not quite - the value is below the blue dot in the precipitation graph. This value is **hail**.

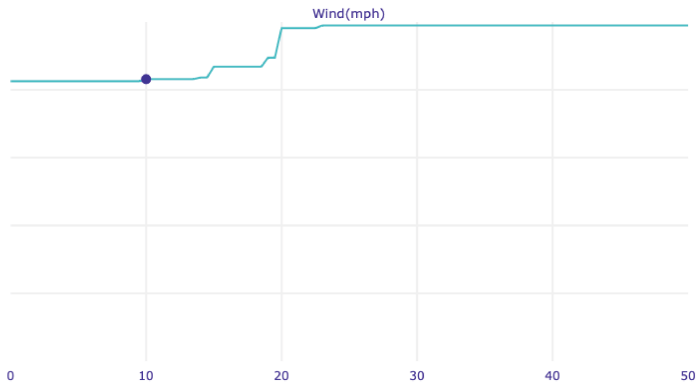


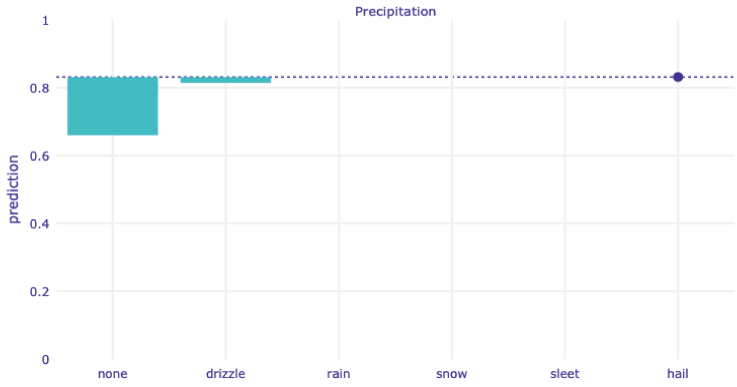
By looking at the explanation image, please select the value for **wind speed** input into the model:

- ○ 20 mph
- ○ 0 mph

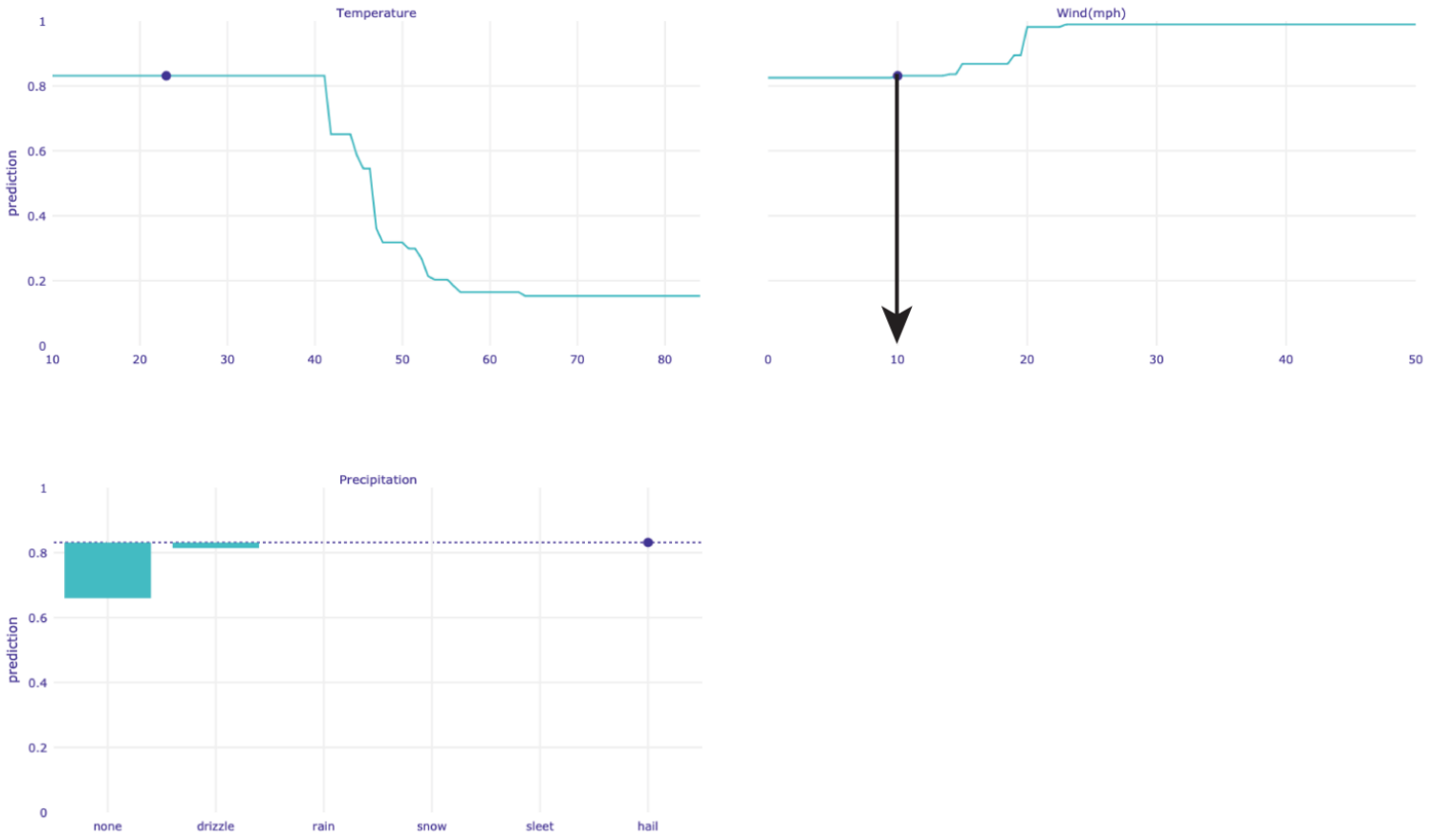
- ○ 10 mph
- ○ 5 mph
- ○ 15 mph



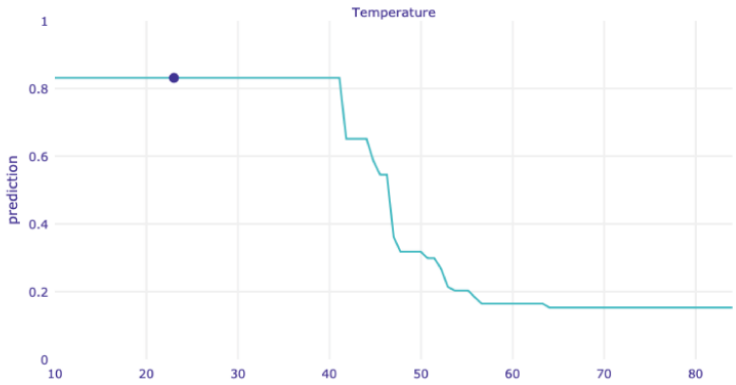


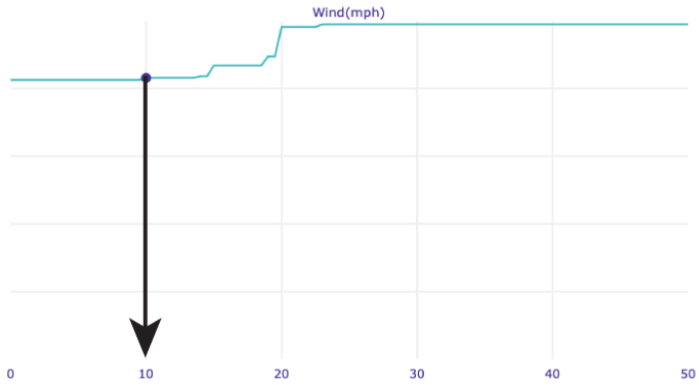


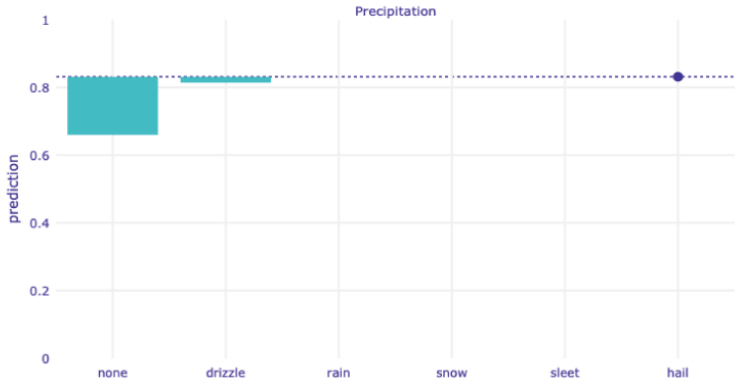
Correct - the value is below the blue dot in the wind(mph) graph. This value is **10 mph**.



Not quite - the value is below the blue dot in the wind(mph) graph. This value is **10 mph**.







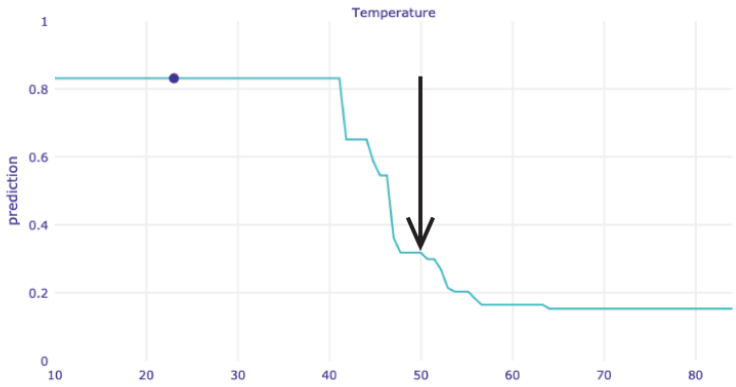
Intro 8

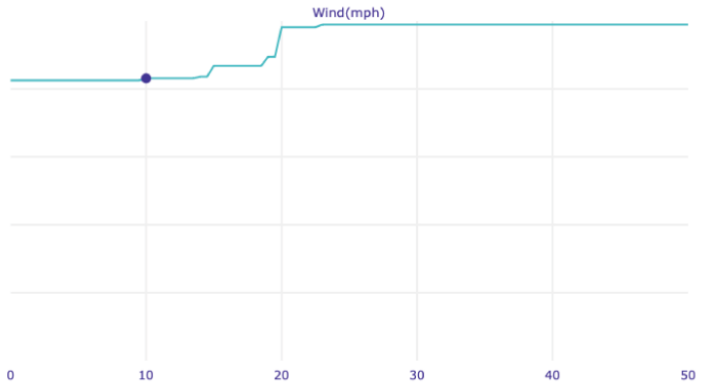
All of the other values on these charts show you what the model's prediction would be, if the value of the corresponding feature was **changed** but the values of all of the other features were **NOT changed.**

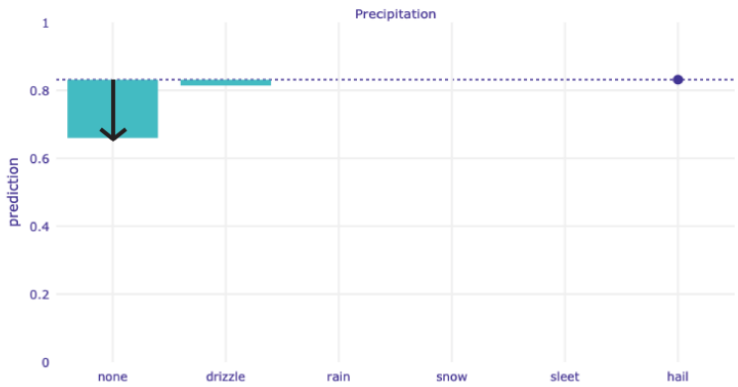
For example, in the Temperature chart below, you can see that if the temperature was **50**, instead of **23**, then the model prediction would be about between **0.3** and **0.4**, meaning that it would now recommend you do NOT put on a coat.

In the Precipitation bar graph below, you can see that if the precipitation was **None**, the model prediction would be between **0.6** and **0.7**. The model would still recommend that you put on a coat, but it would be less certain.

On the other hand, if the value of Temperature was instead 20, or the value of Precipitation was instead sleet, the model prediction would stay the same.





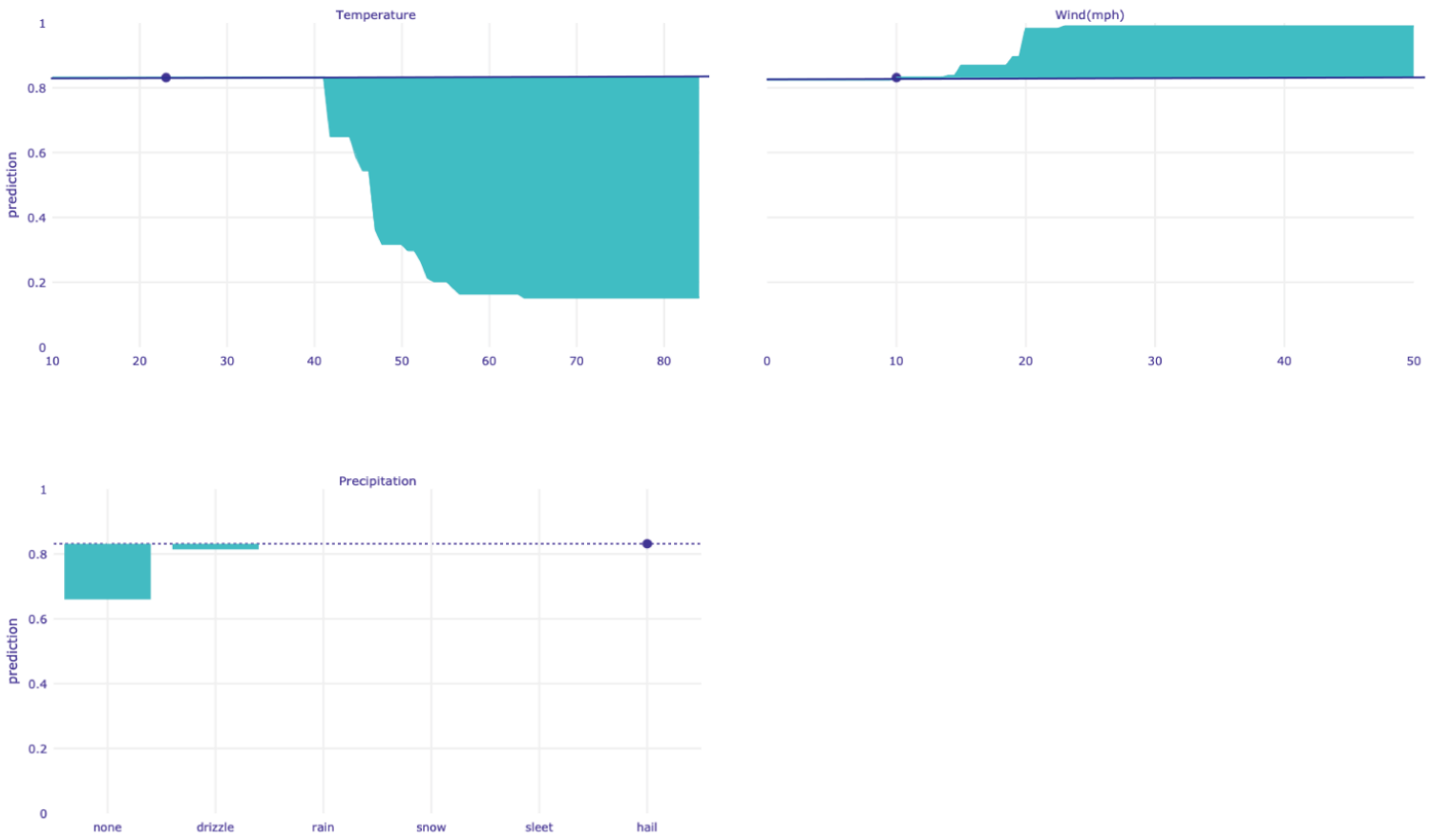


Intro 9

The predictive power of a numerical feature, such as wind speed, depends on how much changing the feature's value will cause the output of the model to change. If a feature is important for this prediction, then changing its value will cause the prediction to change a lot.

You can get an idea of the predictive power of a feature by eyeballing the area between the line that you see, and an imaginary horizontal line at the current prediction value, where the blue dot is.

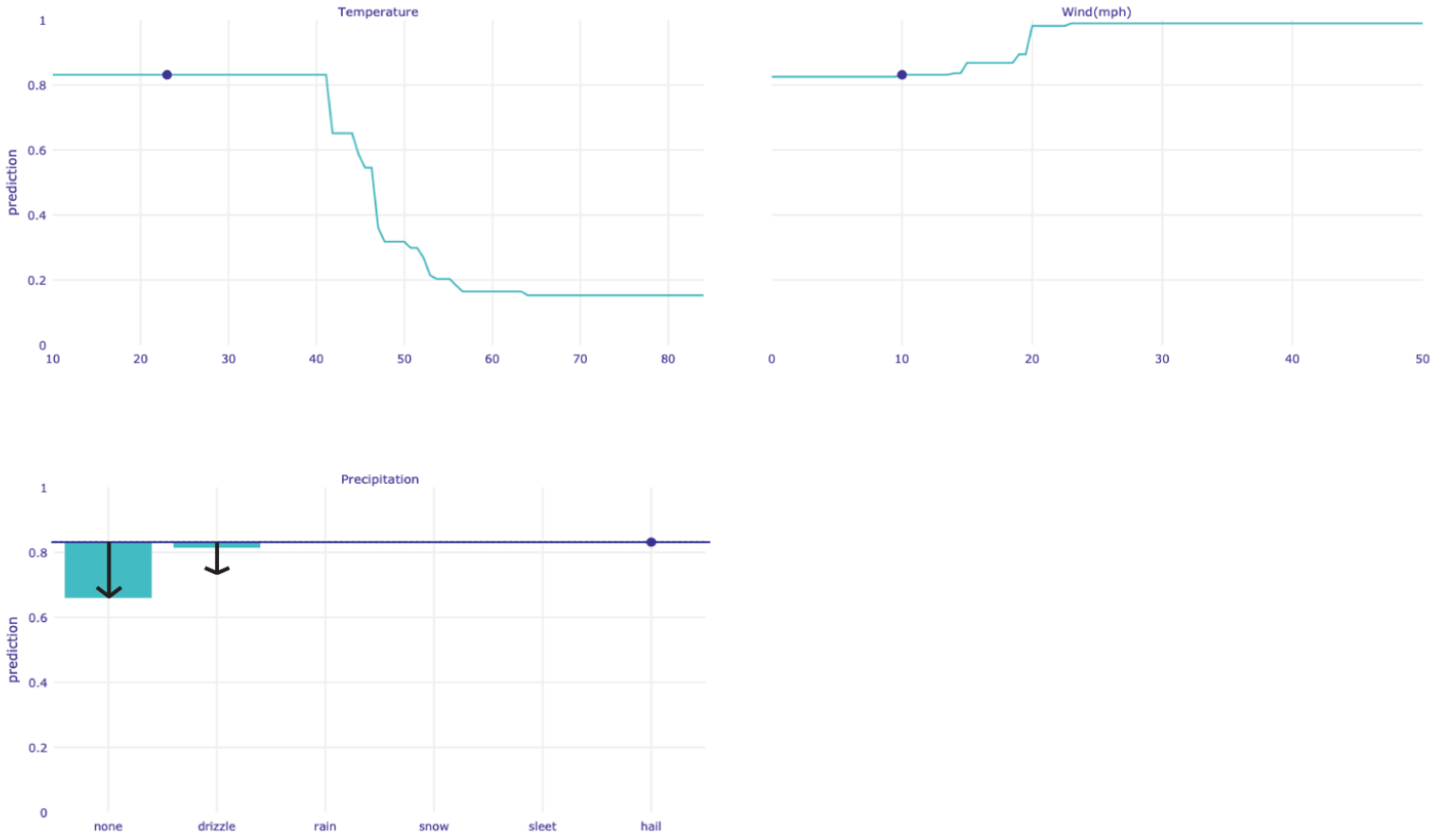
Here, you can see that temperature has more predictive power than wind, because a change in temperature changes the model output more than a change in wind does, and the area between the line and the current prediction value is larger.



Intro 10

The predictive power of a categorical feature is also dependent on how much the prediction outcome changes when the feature value changes.

Except now, instead of looking at the area between two lines, we are eyeballing the sizes of the bars.



Intro 11

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

return 'NO' (do not wear a coat).

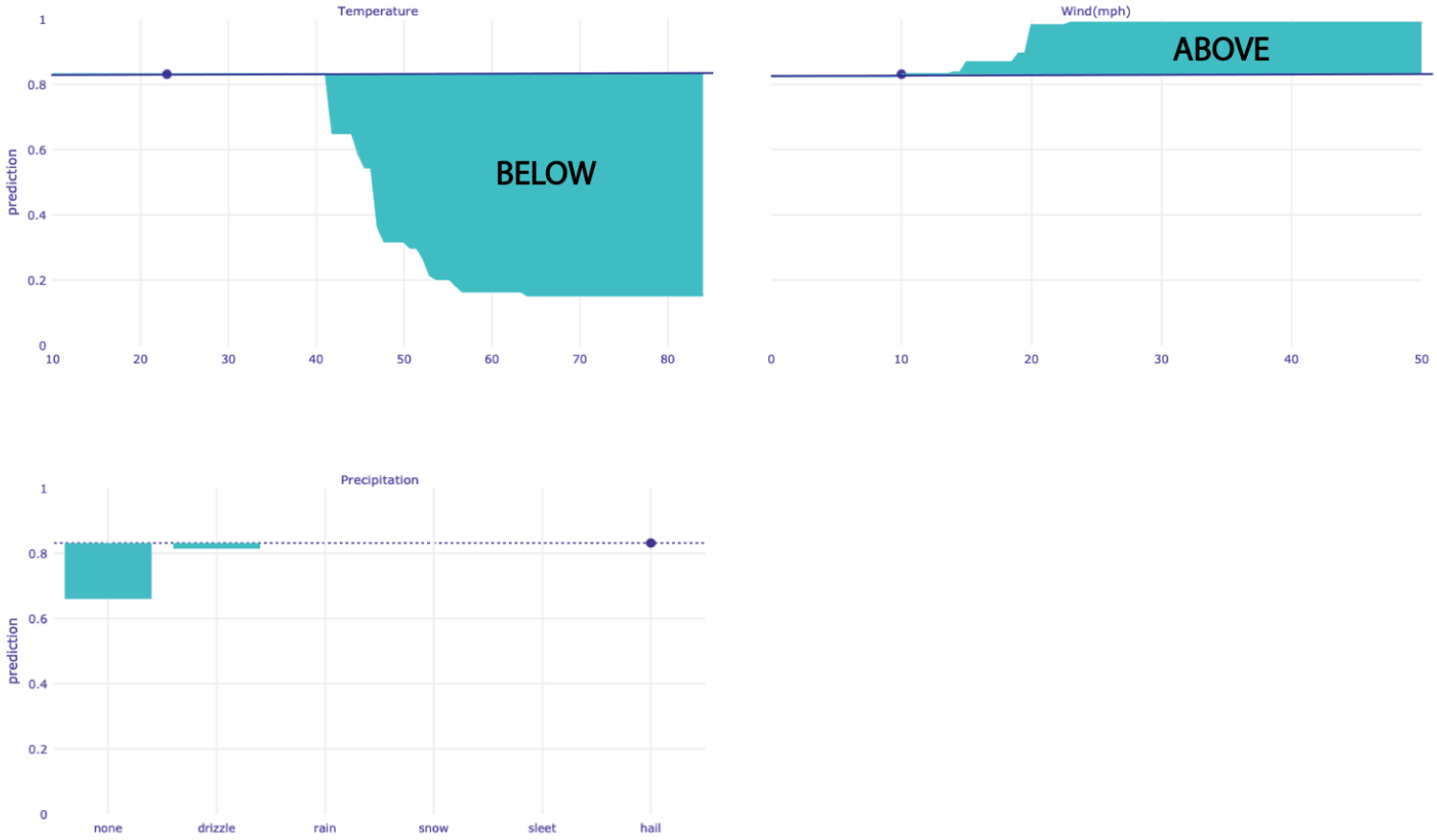
You can tell that a feature value is pushing the model toward predicting 'YES' if changing this value is more likely to **decrease** the model prediction than increase it. For a numerical feature, this means that the area **below** an imaginary horizontal line at the current prediction value dot is **larger** than the area above the line.

For instance, in this example, the area below the current prediction dot in the Temperature chart is large, while the area above the prediction dot is 0. Changing the value of the temperature will probably **decrease** the probability that you should wear a coat. This means the current value of temperature is pushing the model toward predicting 'YES'.

You can tell that a feature value is pushing the model toward predicting 'NO' if changing this value is more likely to **increase** the model prediction than decrease it. For a numerical feature, that means that the area **above** an imaginary horizontal line at the current prediction value is **larger** than the area below the line.

The area above the current prediction dot in the Wind(mph) chart is medium size, while the area below the prediction dot is 0. Changing the value of wind(mph) will probably **increase** the probability that you should wear a coat. This means the current

value of wind(mph) is pushing the model toward predicting 'NO'.



Intro 12

Again, you can tell that a feature value is pushing the model toward predicting 'YES' if changing this value is more likely to

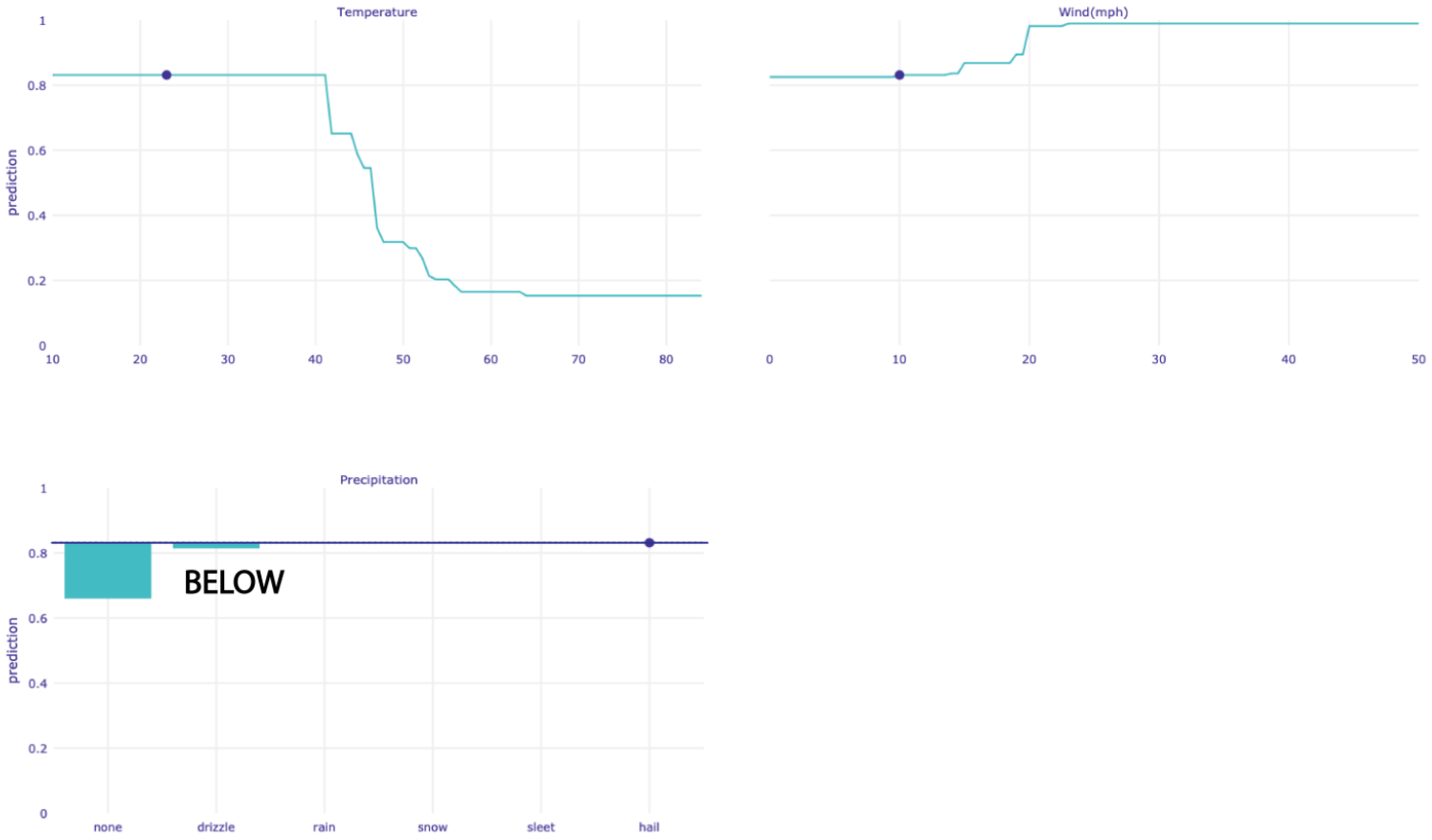
decrease the model prediction than increase it.

For a categorical feature, like wind(mph), this means that the total sum of the bars **below** the blue dot is **larger** than the total sum of the bars above the dot.

And you can tell that a feature is pushing the model toward predicting 'NO' if changing the value is more likely to **increase** the model prediction than decrease it.

For a categorical feature, this means that the total sum of the bars **above** the blue dot is **larger** than the total sum of the bars below the dot.

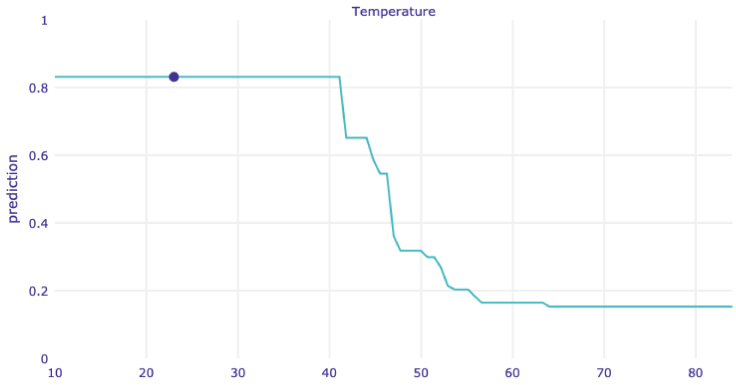
The bars in the Precipitation chart are all pointing down below the blue dot, meaning that changing this feature will probably **decrease** the probability that you should wear a coat. That means that the current value of precipitation is pushing the model toward predicting 'YES'.

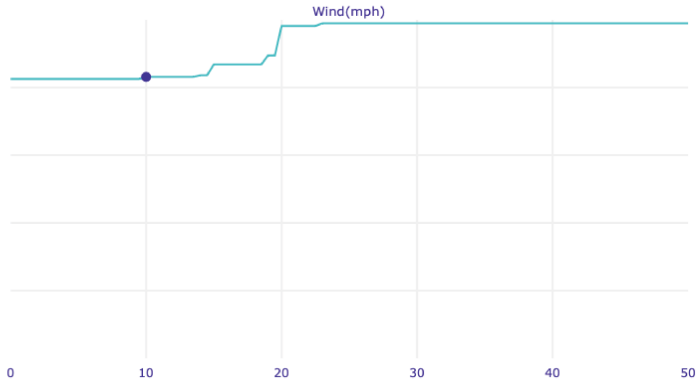


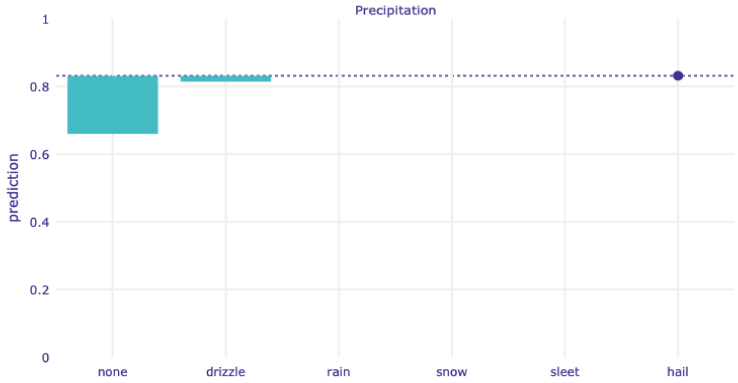
Intro Test 3

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

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







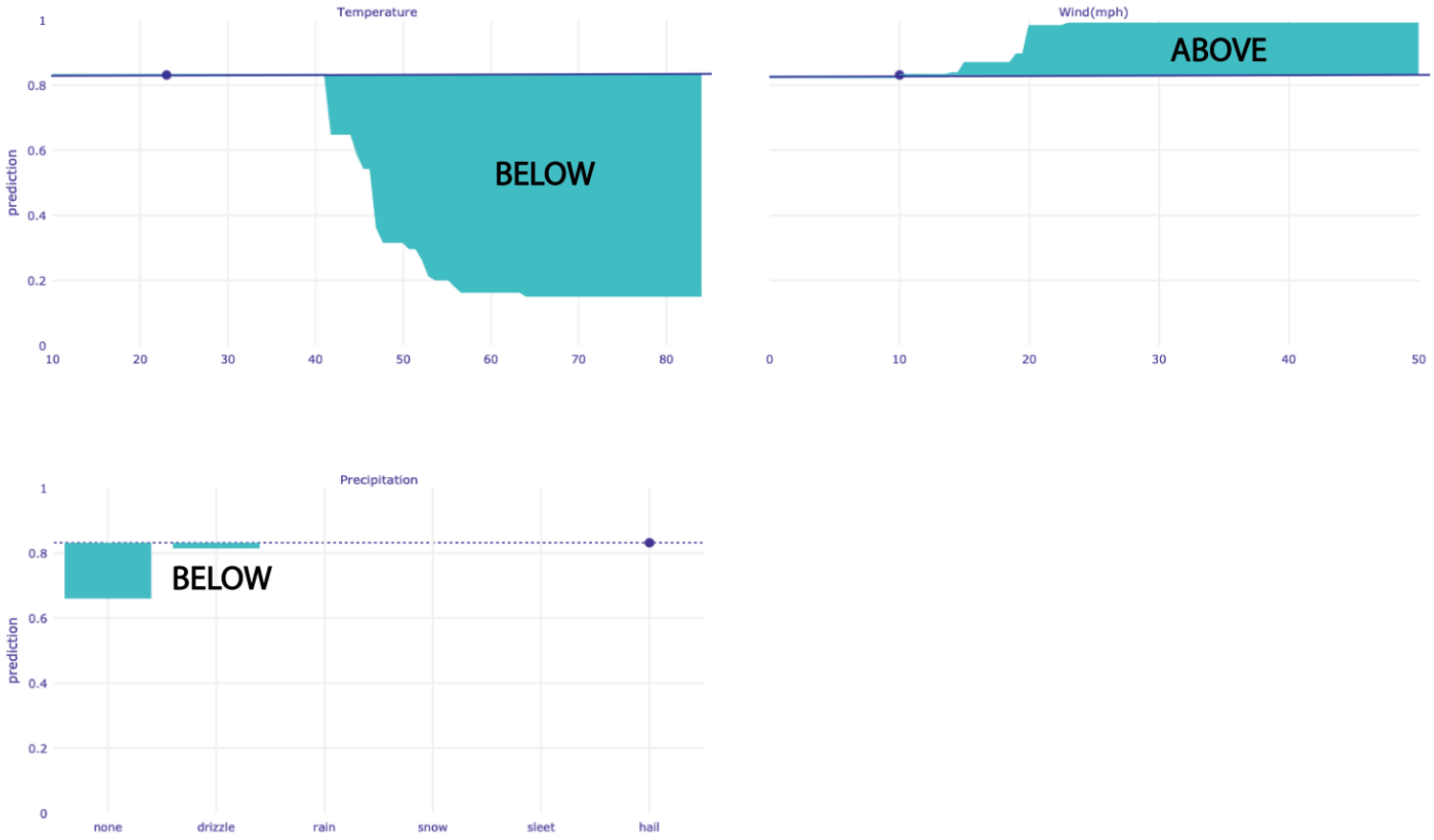
Correct -

In this case, area **below** the current prediction dot in the Temperature chart is large, while the area above the prediction dot is 0. Changing the value of the temperature is probably going to **decrease** the probability that you should wear a coat. So the

current value of temperature is pushing the model toward predicting 'YES'.

All of the bars for the Precipitation chart are pointing down **below** the blue dot. This means that changing precipitation is probably going to **decrease** the probability that you should wear a coat. So the current value of precipitation is also pushing the model toward predicting 'YES'.

Meanwhile, the area **above** the current prediction dot in the Wind(mph) chart is medium size, while the area below the prediction dot is 0. Changing the value of wind(mph) is probably going to **increase** the probability that you should wear a coat. So the current value of wind is pushing the model toward predicting 'NO'.

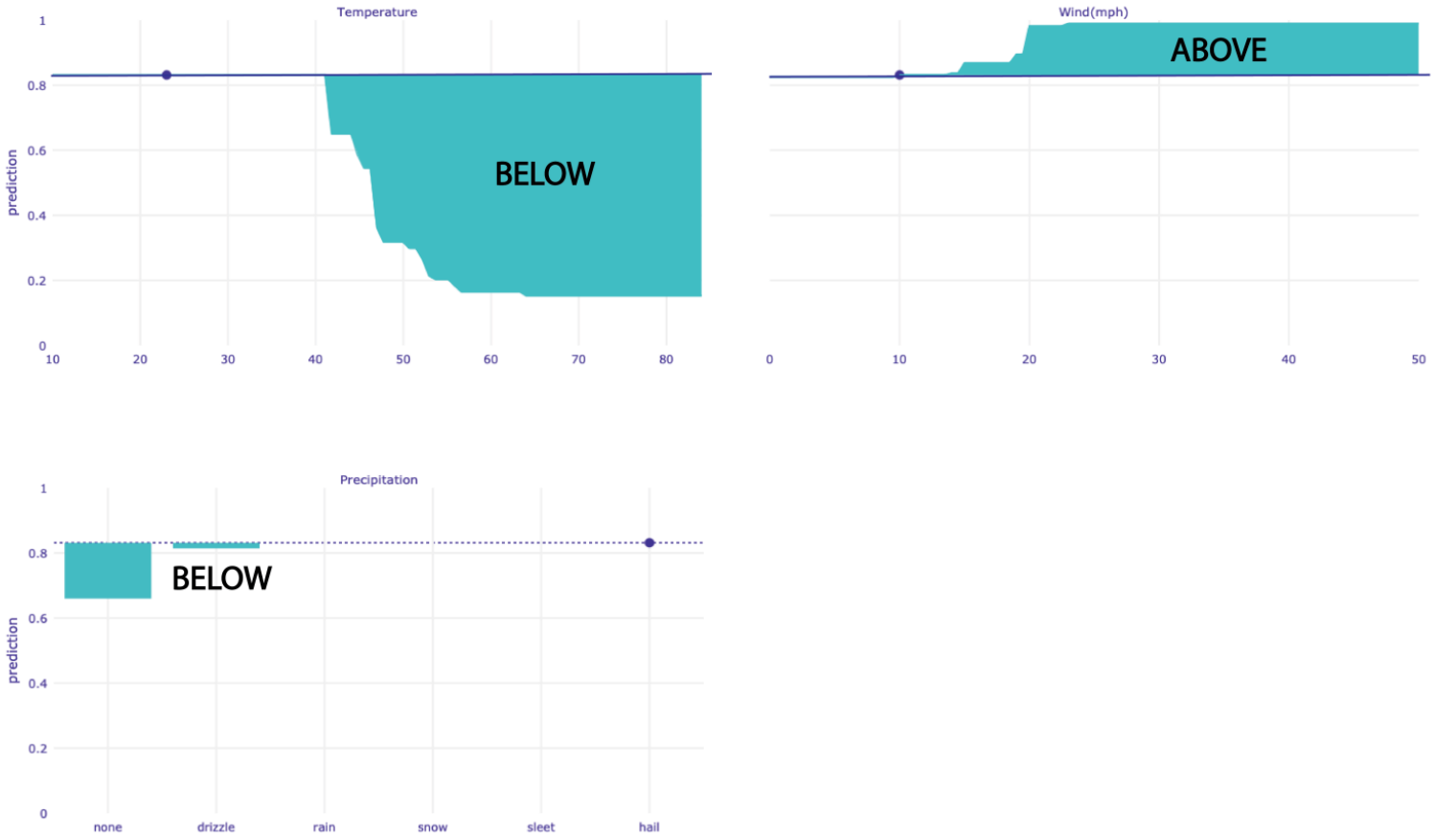


Not quite -

In this case, area **below** the current prediction dot in the Temperature chart is large, while the area above the prediction dot is 0. Changing the value of the temperature is probably going to **decrease** the probability that you should wear a coat. So the current value of temperature is pushing the model toward predicting 'YES'.

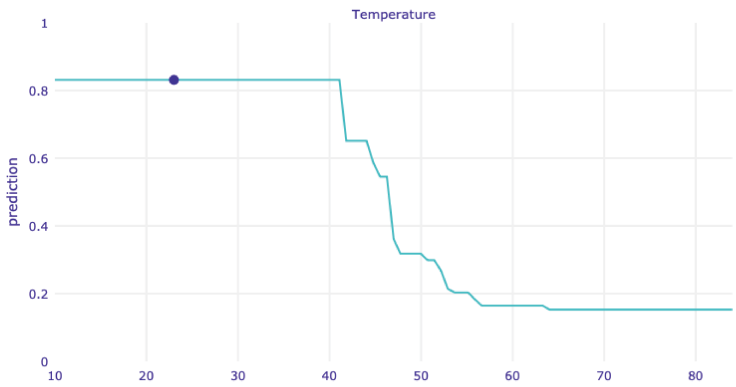
All of the bars for the Precipitation chart are pointing down **below** the blue dot. This means that changing precipitation is probably going to **decrease** the probability that you should wear a coat. So the current value of precipitation is also pushing the model toward predicting 'YES'.

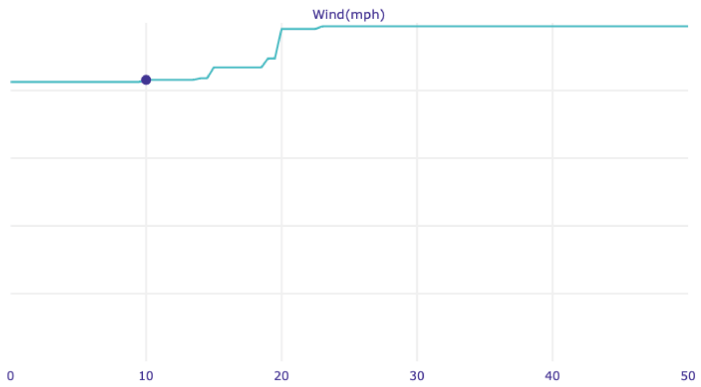
Meanwhile, the area **above** the current prediction dot in the Wind(mph) chart is medium size, while the area below the prediction dot is 0. Changing the value of wind(mph) is probably going to **increase** the probability that you should wear a coat. So the current value of wind is pushing the model toward predicting 'NO'.

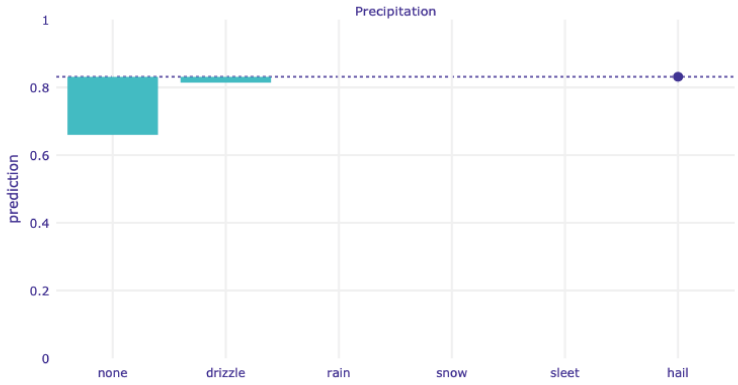


By looking at the explanation image, which factor(s) are pushing the model toward predicting 'NO'?

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







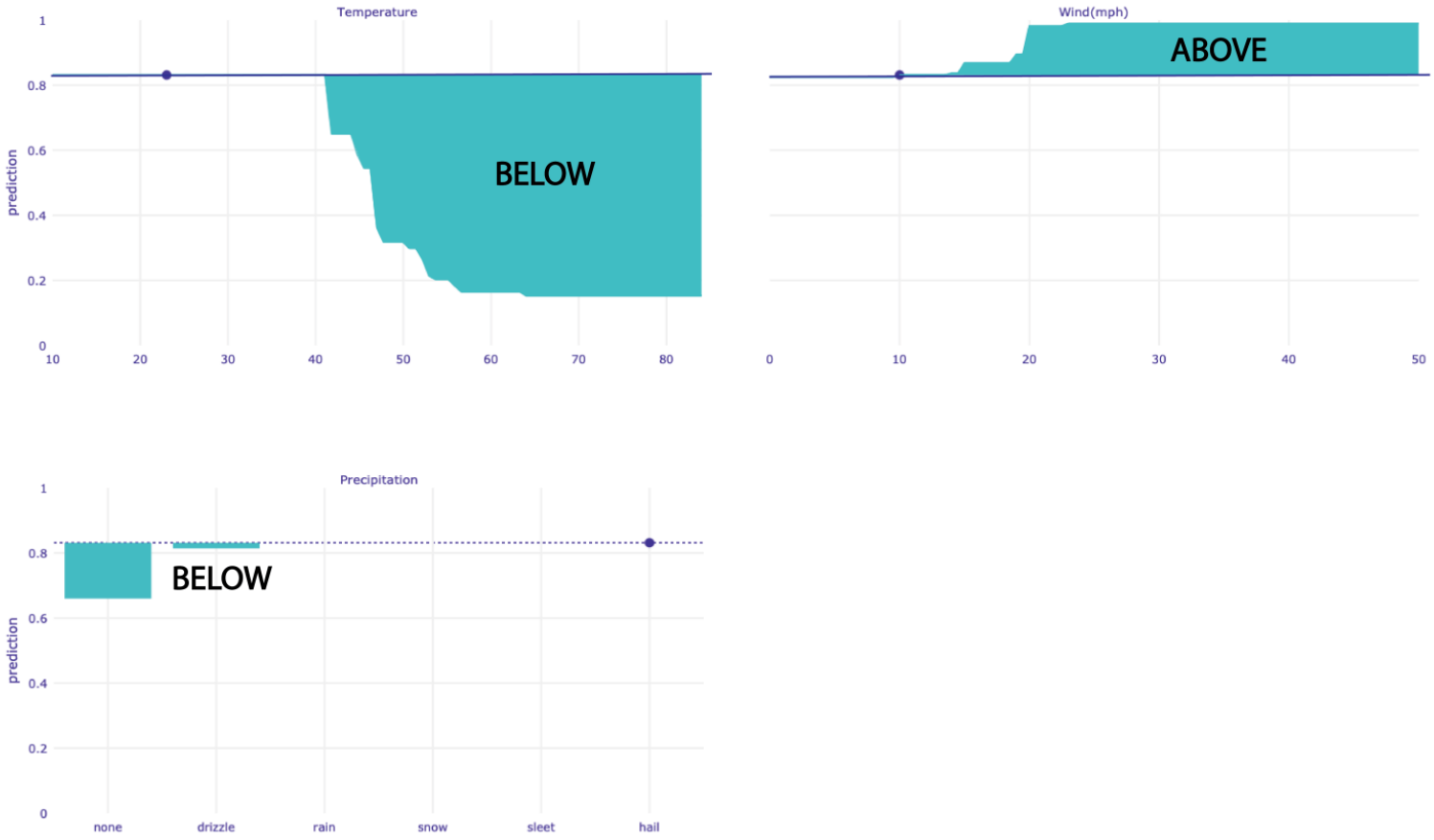
Correct -

In this case, area **below** the current prediction dot in the Temperature chart is large, while the area above the prediction dot is 0. Changing the value of the temperature is probably going to **decrease** the probability that you should wear a coat. So the current value of temperature is pushing the model toward

predicting 'YES'.

All of the bars for the Precipitation chart are pointing down **below** the blue dot. This means that changing precipitation is probably going to **decrease** the probability that you should wear a coat. So the current value of precipitation is also pushing the model toward predicting 'YES'.

Meanwhile, the area **above** the current prediction dot in the Wind(mph) chart is medium size, while the area below the prediction dot is 0. Changing the value of wind(mph) is probably going to **increase** the probability that you should wear a coat. So the current value of wind is pushing the model toward predicting 'NO'.

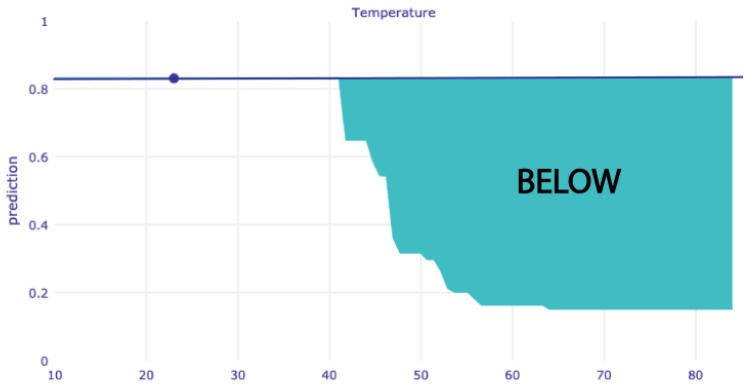


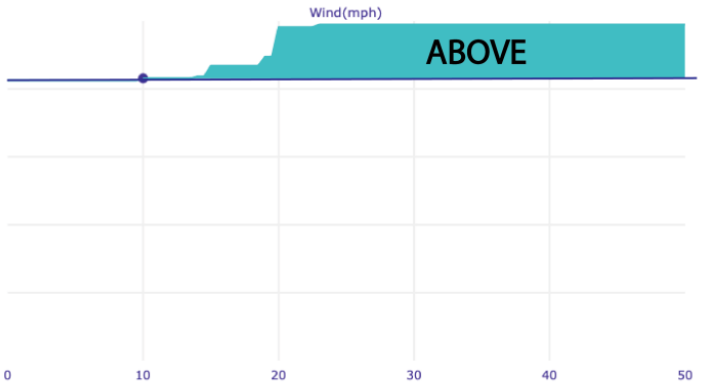
Not quite -

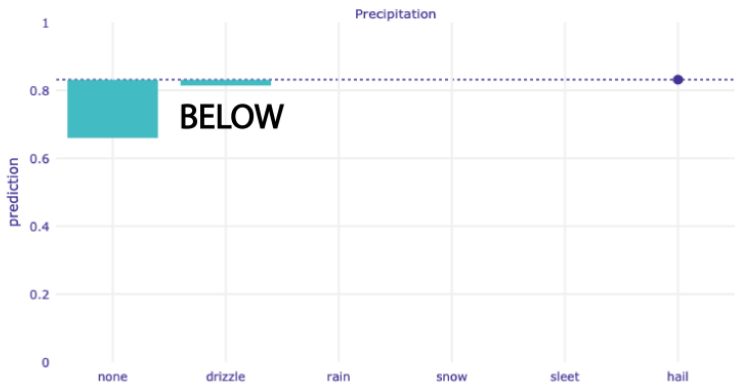
In this case, area **below** the current prediction dot in the Temperature chart is large, while the area above the prediction dot is 0. Changing the value of the temperature is probably going to **decrease** the probability that you should wear a coat. So the current value of temperature is pushing the model toward predicting 'YES'.

All of the bars for the Precipitation chart are pointing down **below** the blue dot. This means that changing precipitation is probably going to **decrease** the probability that you should wear a coat. So the current value of precipitation is also pushing the model toward predicting 'YES'.

Meanwhile, the area **above** the current prediction dot in the Wind(mph) chart is medium size, while the area below the prediction dot is 0. Changing the value of wind(mph) is probably going to **increase** the probability that you should wear a coat. So the current value of wind is pushing the model toward predicting 'NO'.

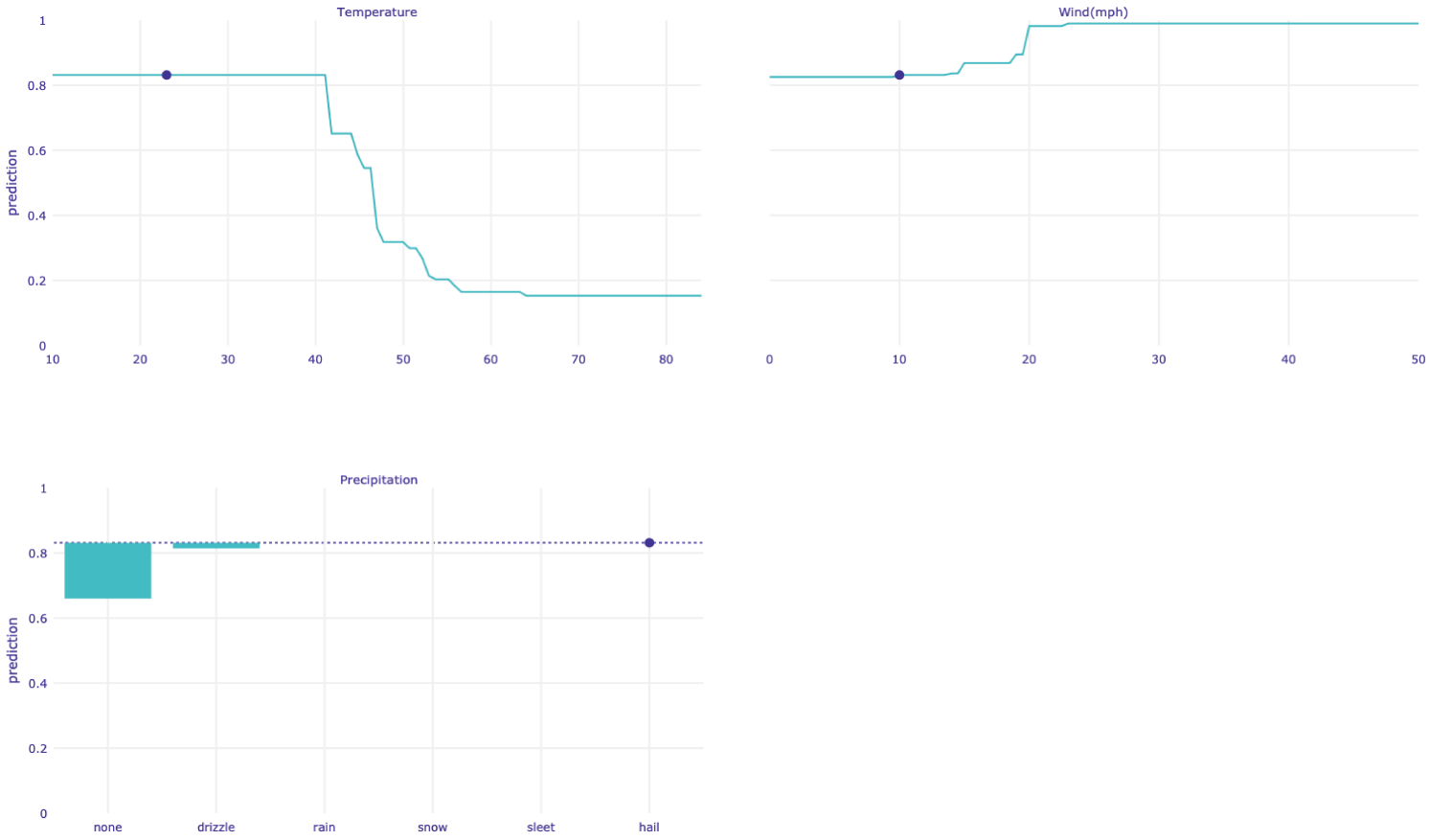




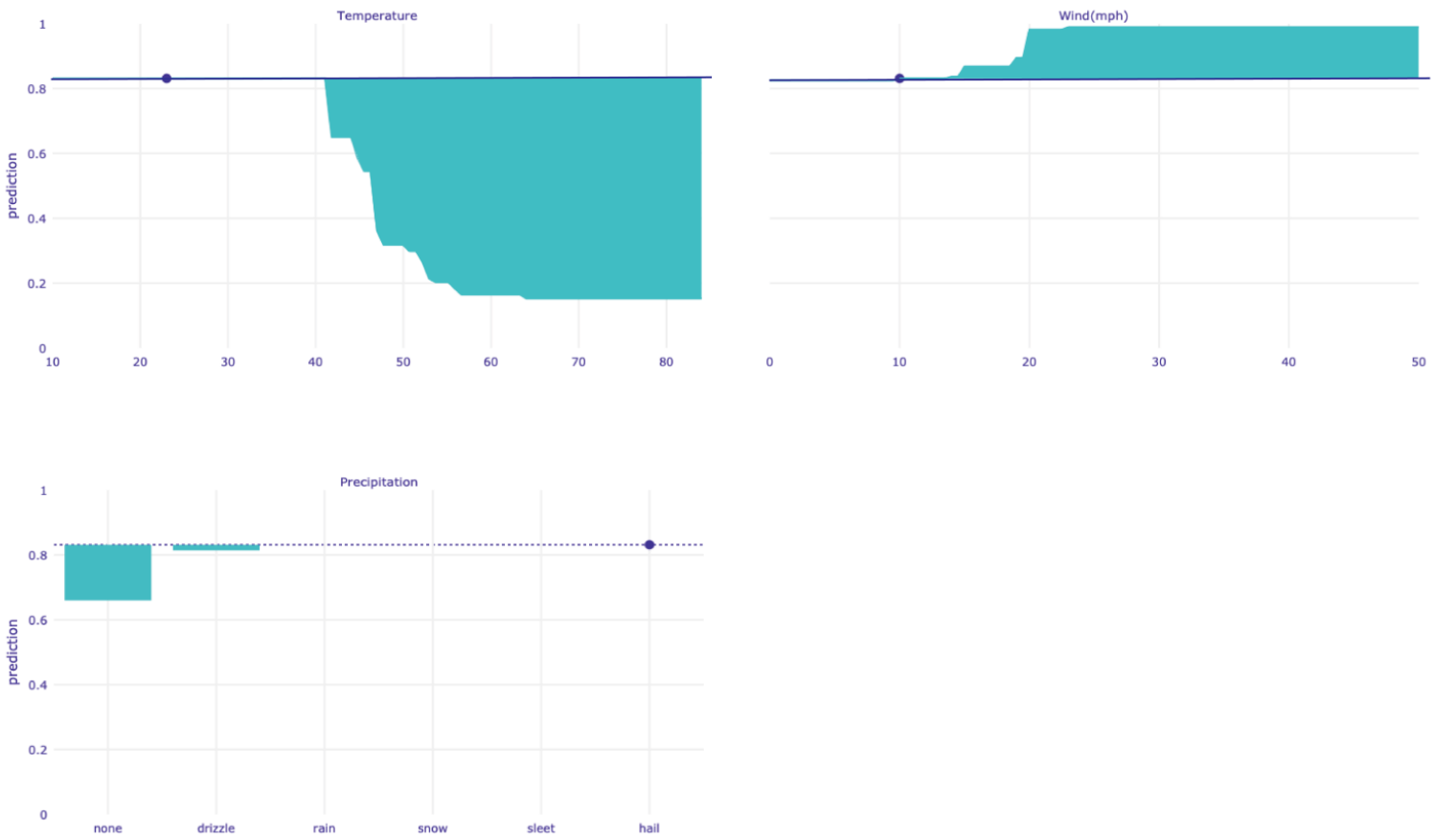


Which factor has the greatest predictive power?

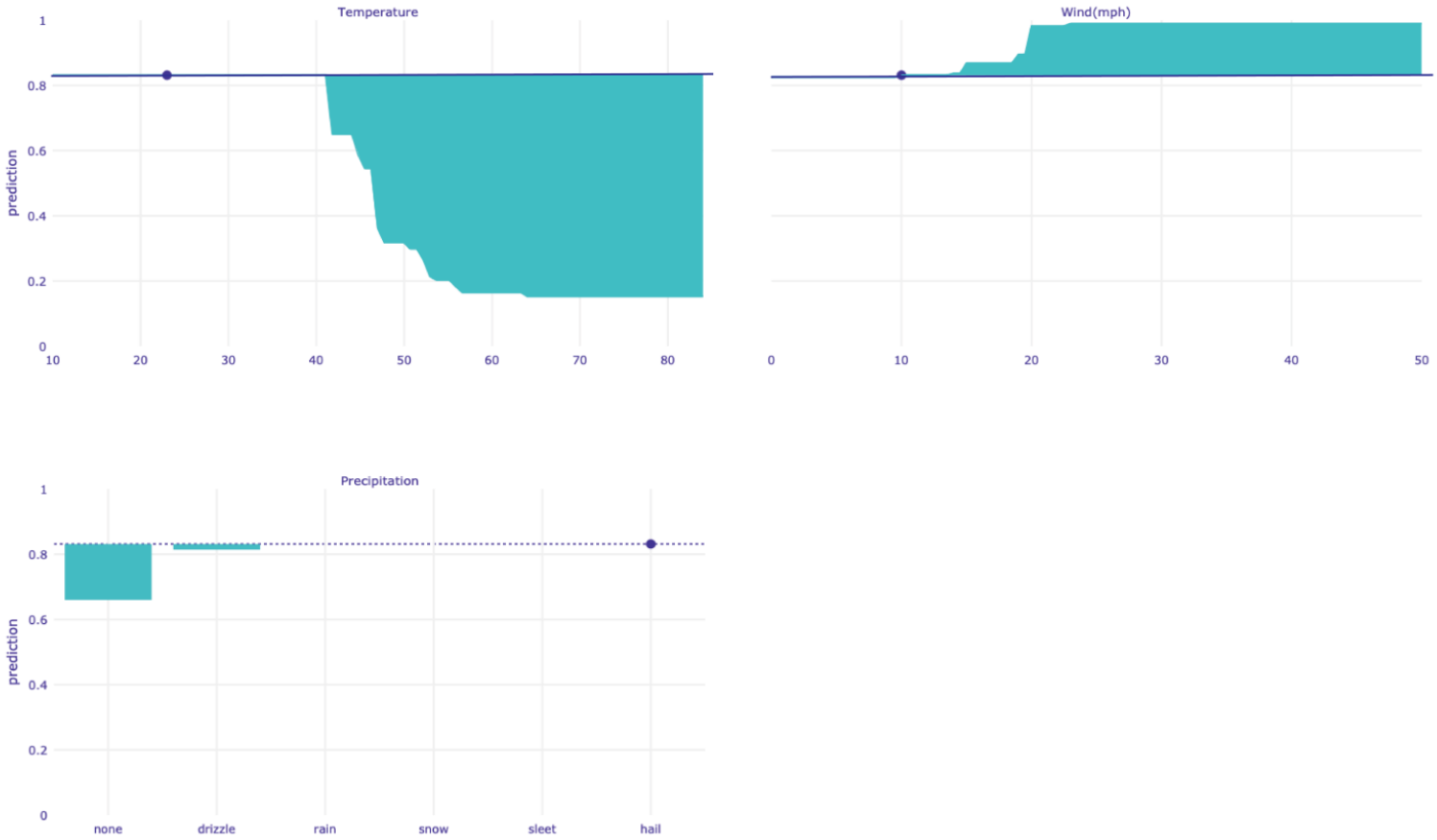
- ○ Temperature
- ○ Wind
- ○ Precipitation



Correct - In this case, the area between the line and the current prediction in the Temperature chart is larger than the area between the line and the current prediction in the Wind(mph) chart, and the bars in the Precipitation chart. Changing the Temperature can cause a larger change in model output than changing the Wind(mph) or the Precipitation.



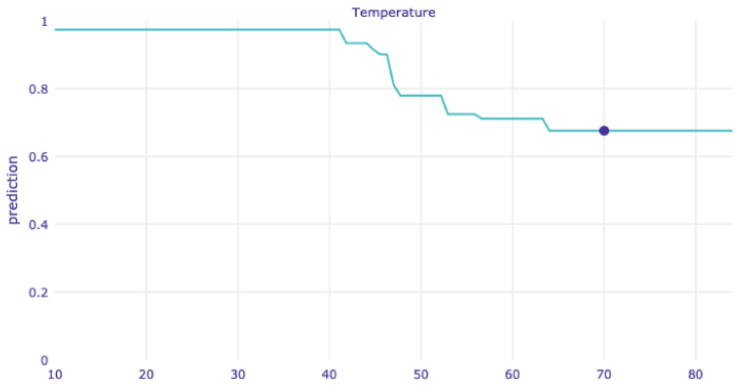
Not quite - In this case, the area between the line and the current prediction in the Temperature chart is larger than the area between the line and the current prediction in the Wind(mph) chart, and the bars in the Precipitation chart. Changing the Temperature can cause a larger change in model output than changing the Wind(mph) or the Precipitation.

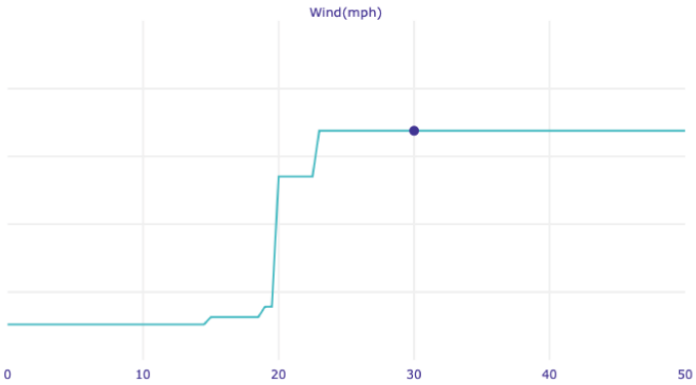


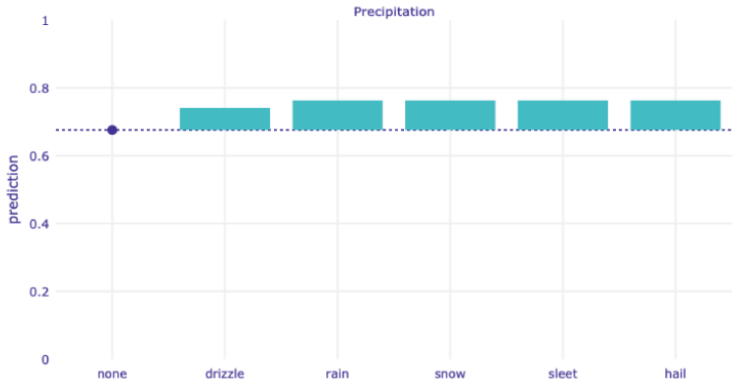
Intro Test 4

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

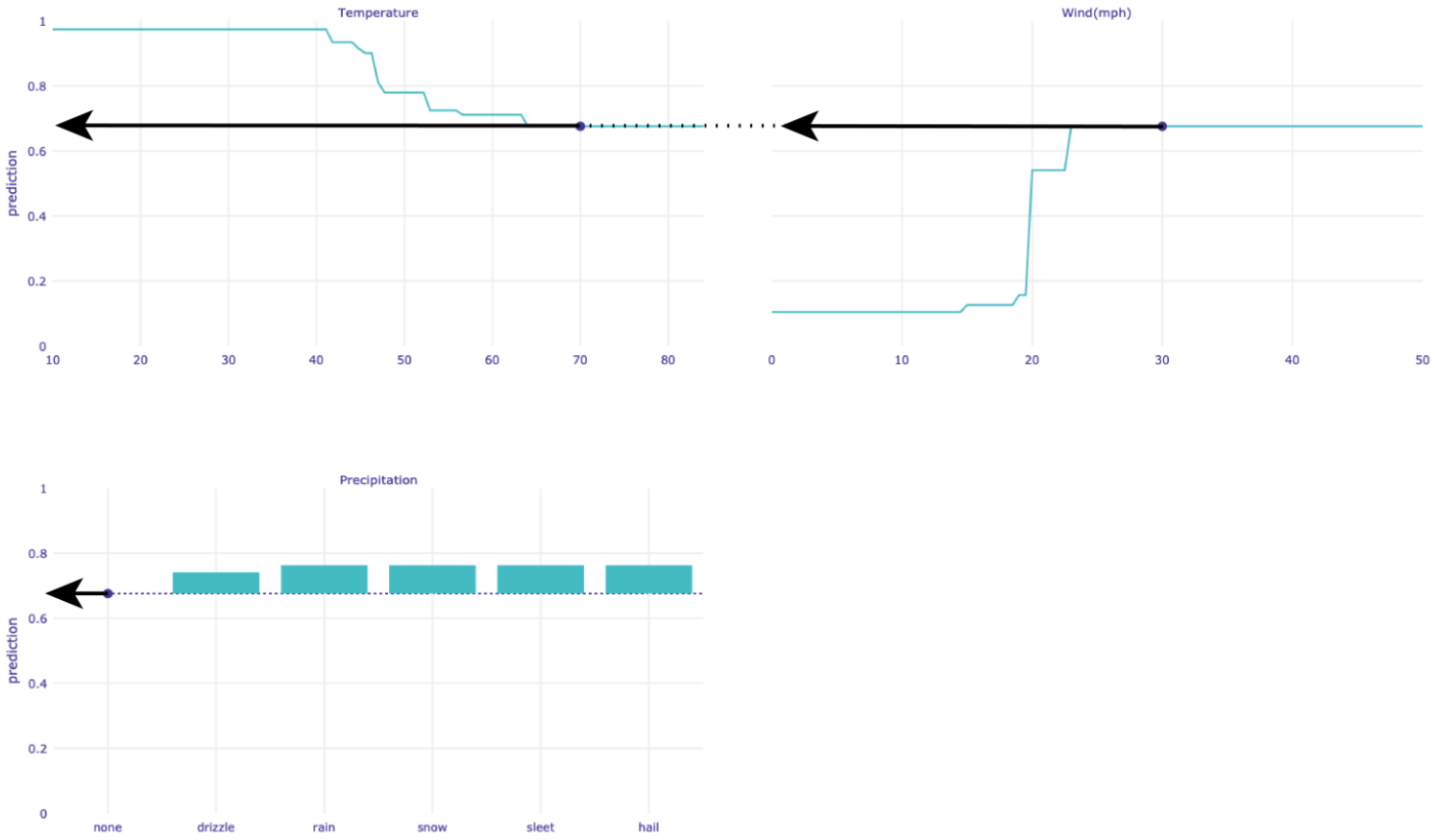
- ○ YES, you should wear a coat
- ○ NO, do not wear a coat



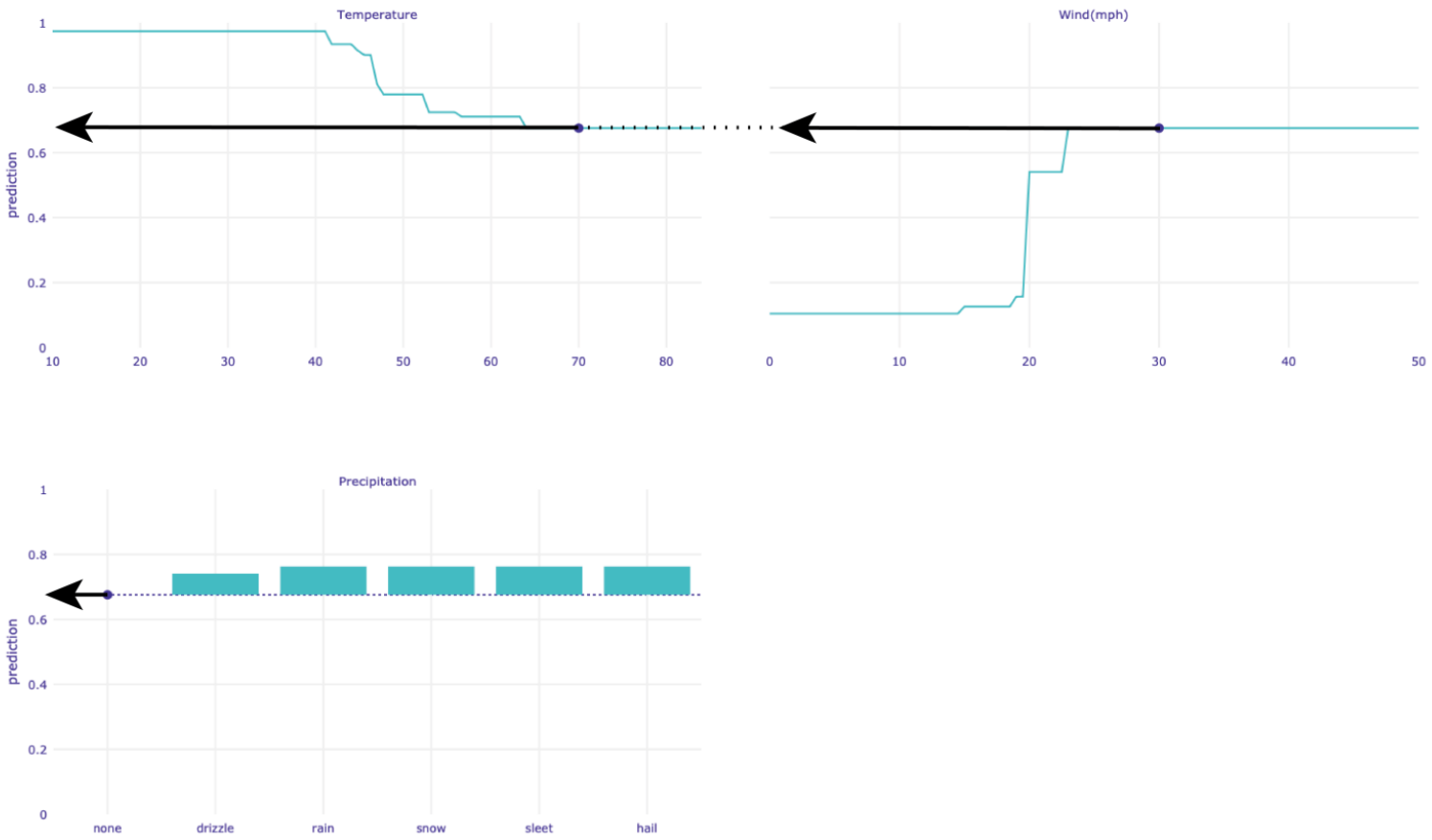




Correct. In this case, the model prediction is 0.634, which is greater than 0.5, so the model will return 'YES'.

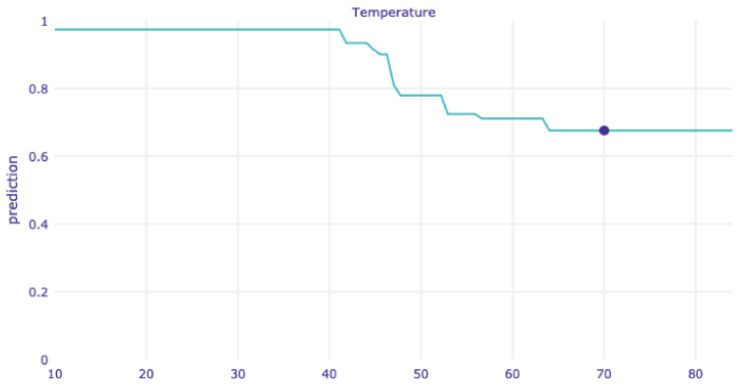


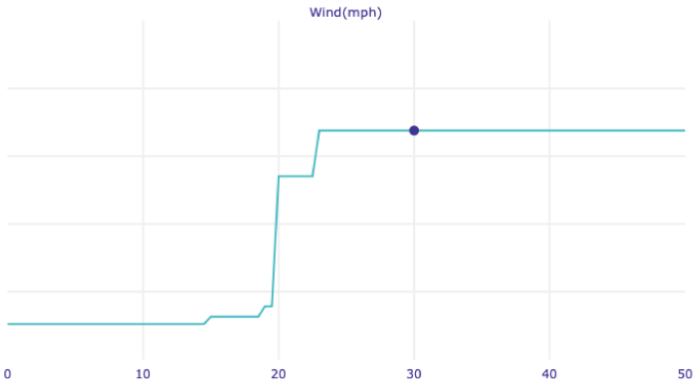
Incorrect. In this case, the model prediction is 0.634, which is greater than 0.5, so the model will return 'YES'.

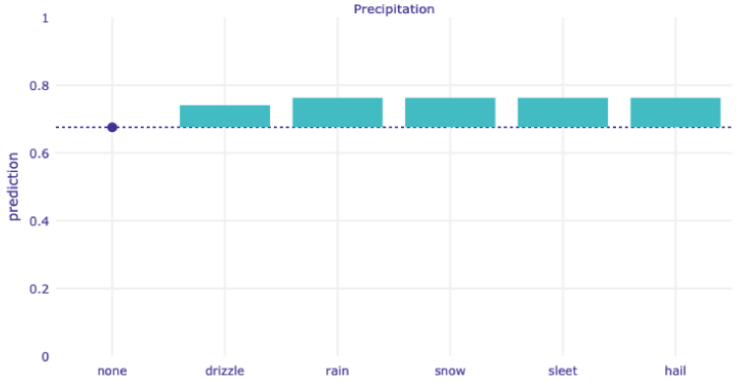


By looking at the explanation image, please select the value for **temperature** input into the model:

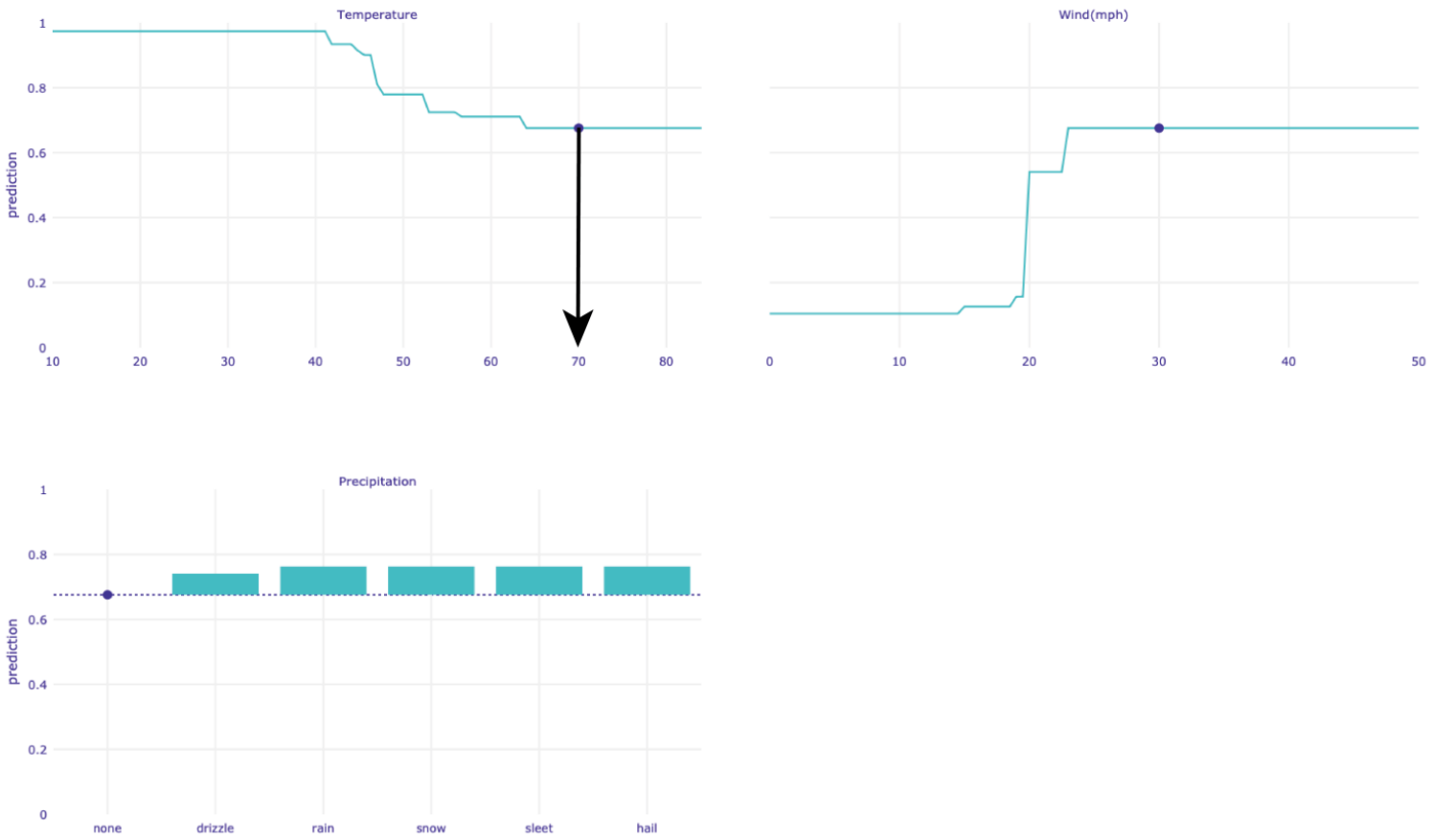
- ○ 84
- ○ 70
- ○ 61
- ○ 56
- ○ 37



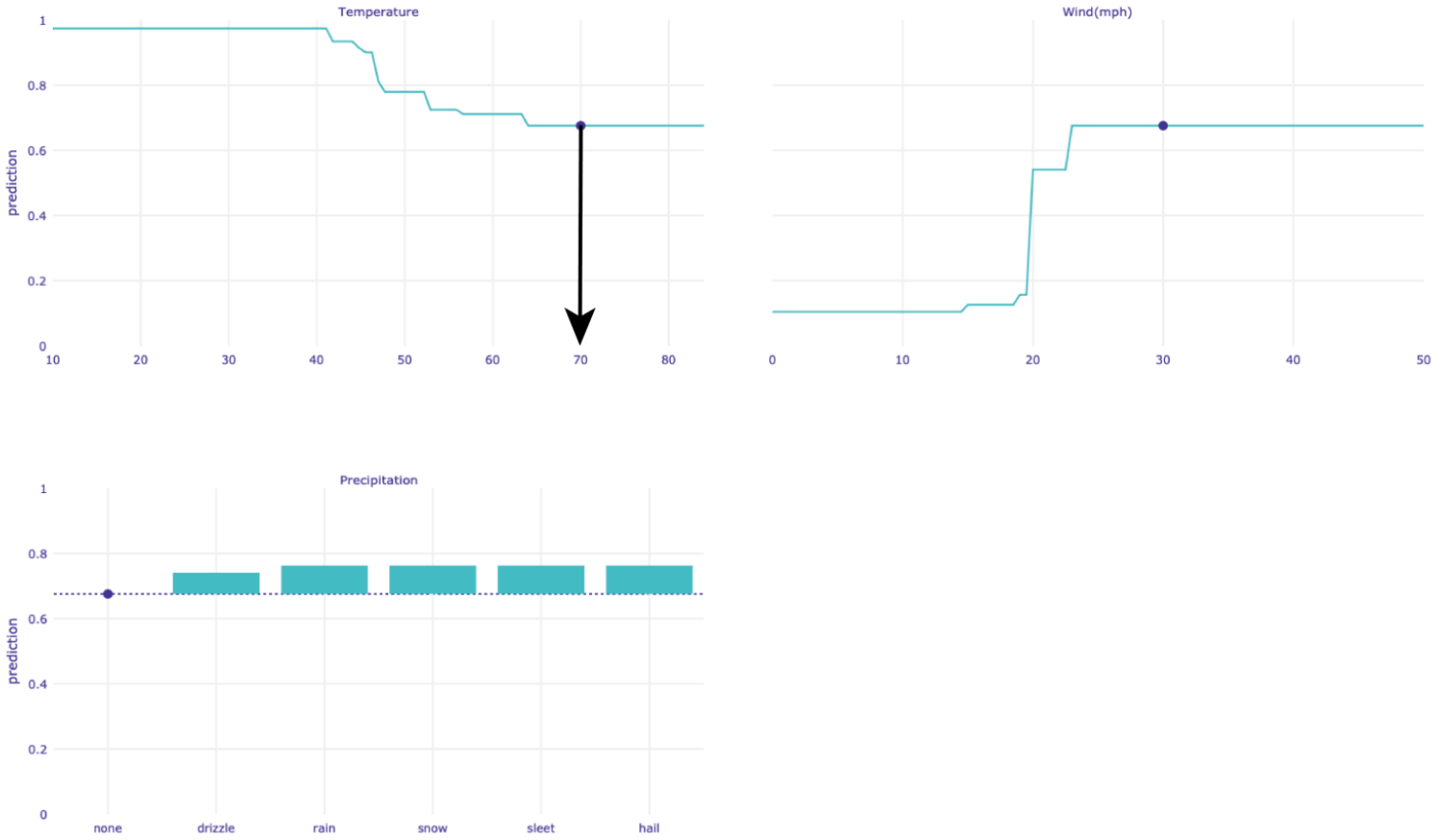




Correct - the value is under the **blue dot** in the Temperature chart. This value is 70.

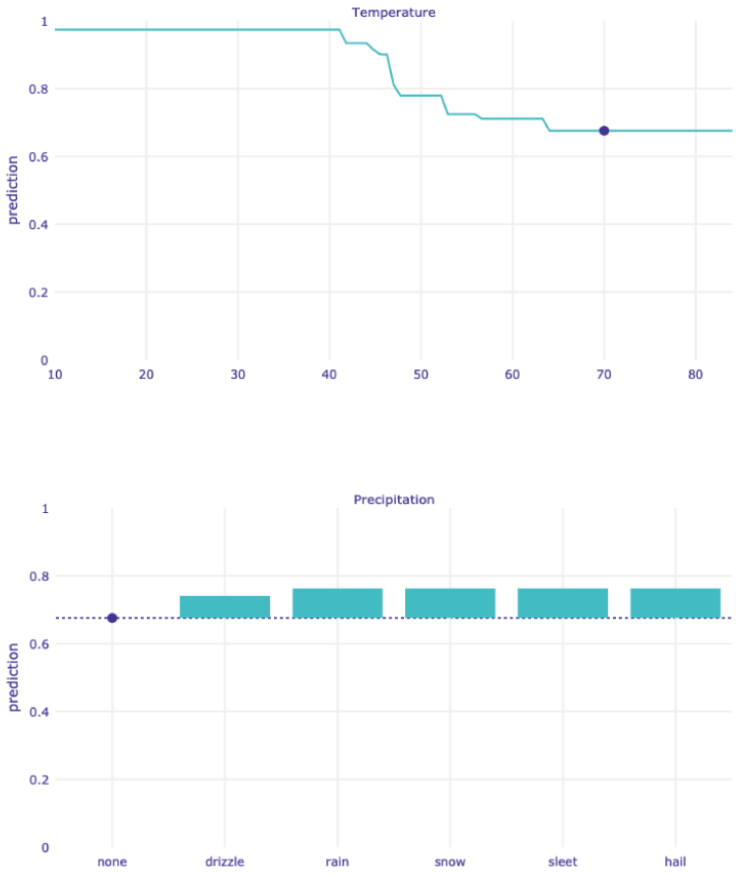


Incorrect - the value is under the **blue dot** in the Temperature chart. This value is 70.



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

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

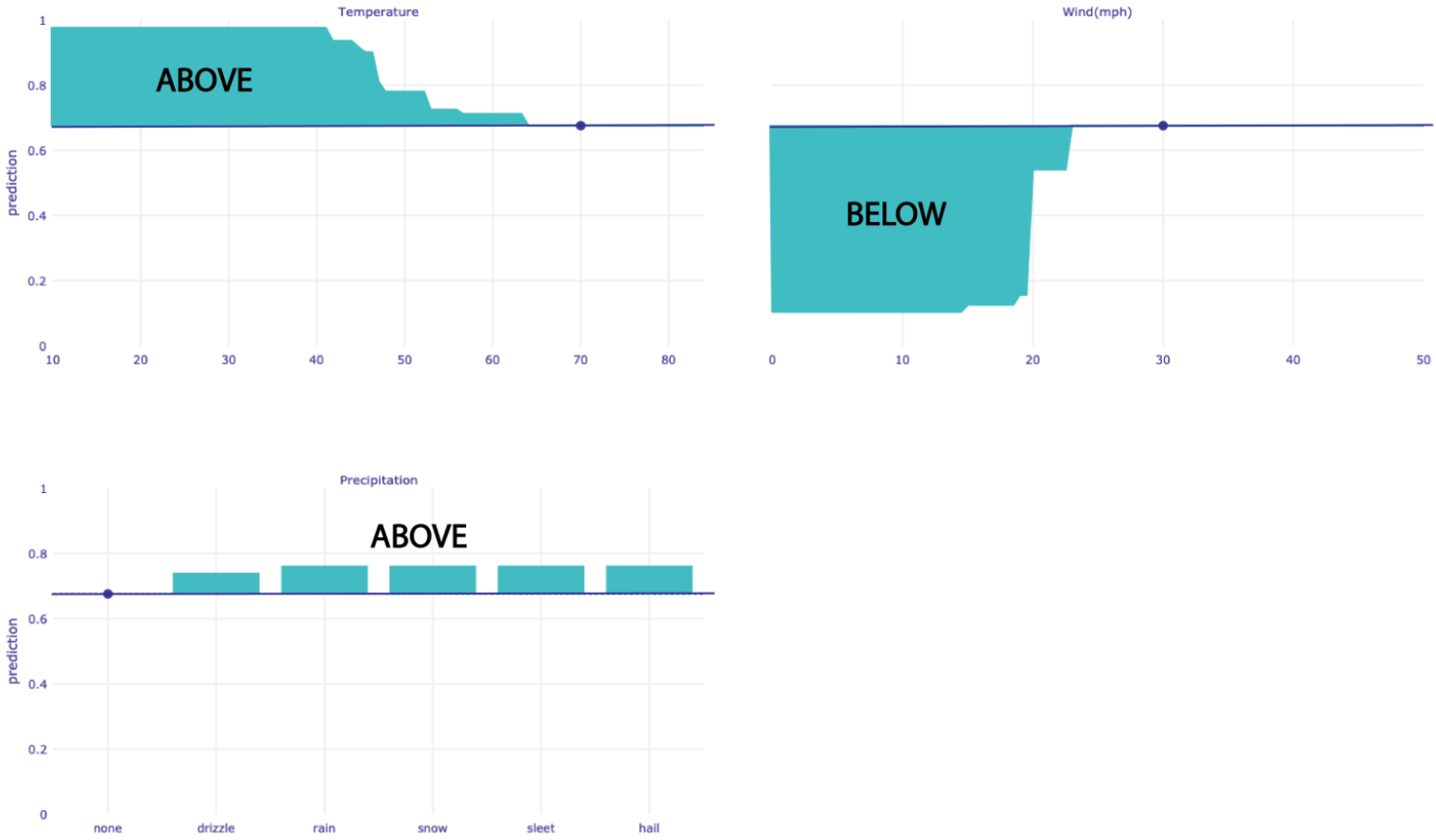


Correct -

In this case, area **above** the current prediction dot in the Temperature chart is large, while the area below the prediction dot is 0. Changing the value of the temperature is probably going to **increase** the probability that you should wear a coat. So the current value of temperature is pushing the model toward predicting 'NO'.

All of the bars for the Precipitation chart are pointing up **above** the prediction dot. This means that changing precipitation is probably going to **increase** the probability that you should wear a coat. So the current value of precipitation is also pushing the model toward predicting 'NO'.

Meanwhile, the area **below** the current prediction line in the Wind(mph) chart is large, while the area above the prediction line is 0. Changing the value of wind(mph) is probably going to **decrease** the probability that you should wear a coat. So the current value of wind is pushing the model toward predicting 'YES'.

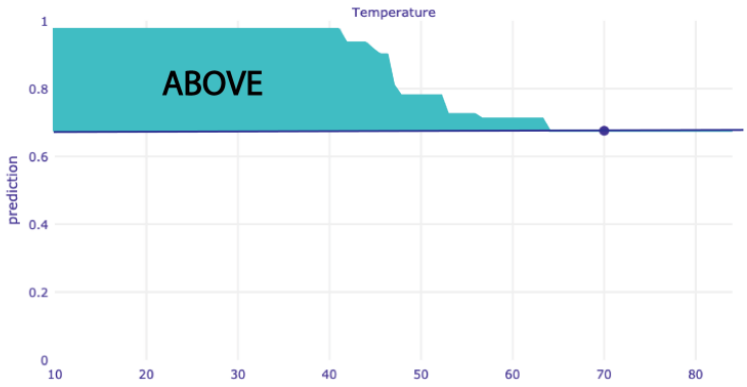


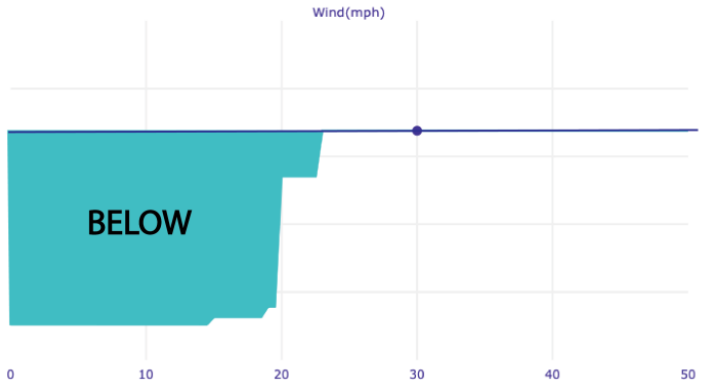
Not quite -

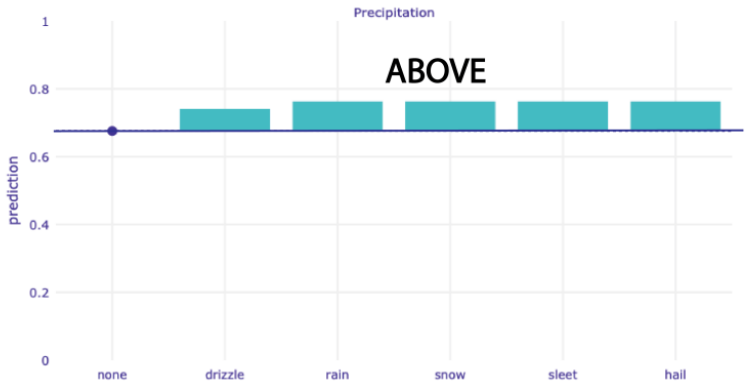
In this case, area **above** the current prediction dot in the Temperature chart is large, while the area below the prediction dot is 0. Changing the value of the temperature is probably going to **increase** the probability that you should wear a coat. So the current value of temperature is pushing the model toward predicting 'NO'.

All of the bars for the Precipitation chart are pointing up **above** the prediction dot. This means that changing precipitation is probably going to **increase** the probability that you should wear a coat. So the current value of precipitation is also pushing the model toward predicting 'NO'.

Meanwhile, the area **below** the current prediction line in the Wind(mph) chart is large, while the area above the prediction line is 0. Changing the value of wind(mph) is probably going to **decrease** the probability that you should wear a coat. So the current value of wind is pushing the model toward predicting 'YES'.

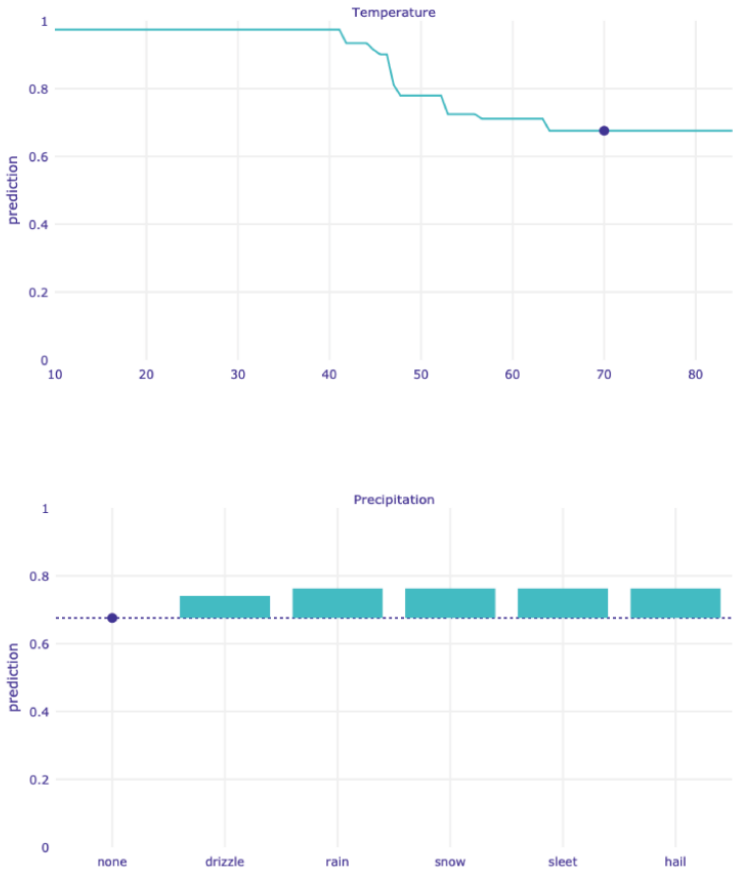




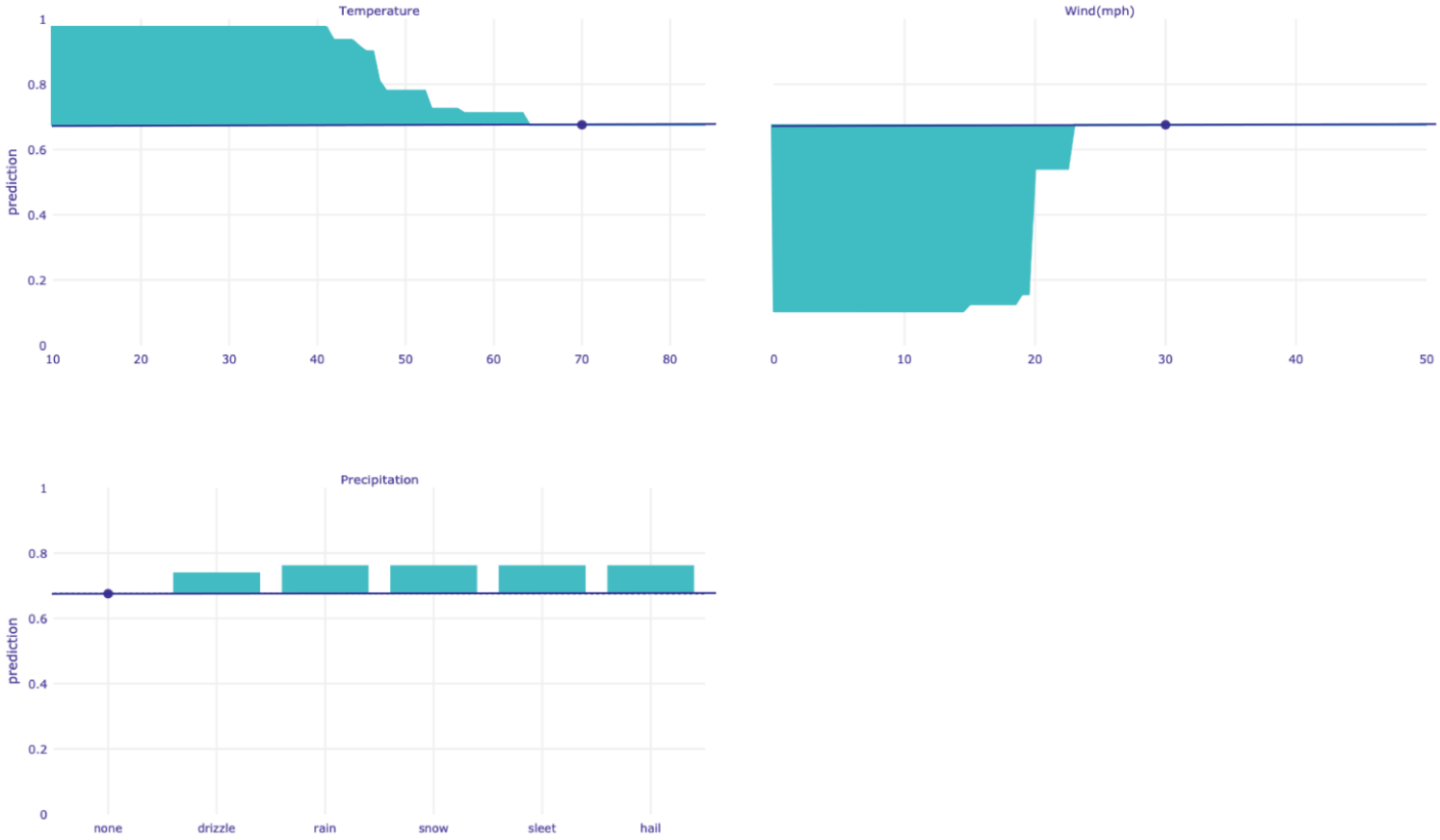


Which factor has the greatest predictive power?

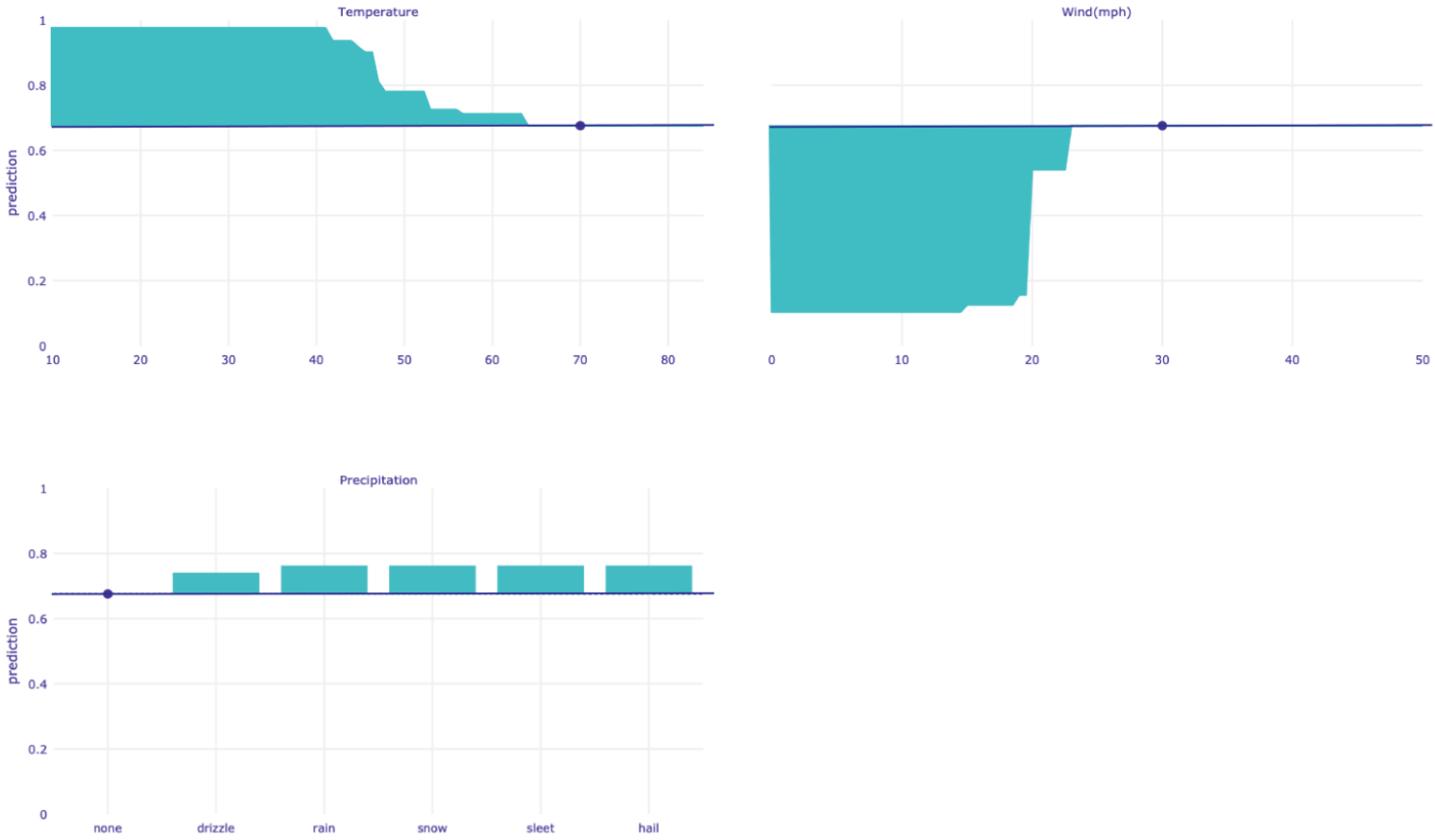
- ○ Temperature
- ○ Wind
- ○ Precipitation



Correct - In this case, the area between the line and the current prediction in the Wind(mph) chart is larger than the area between the line and the current prediction in the Temperature chart, and the bars in the Precipitation chart. Changing the wind(mph) can cause a larger change in model output than changing the Temperature or the Precipitation.



Not quite - In this case, the area between the line and the current prediction in the Wind(mph) chart is larger than the area between the line and the current prediction in the Temperature chart, and the bars in the Precipitation chart. Changing the wind(mph) can cause a larger change in model output than changing the Temperature or the Precipitation.



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.

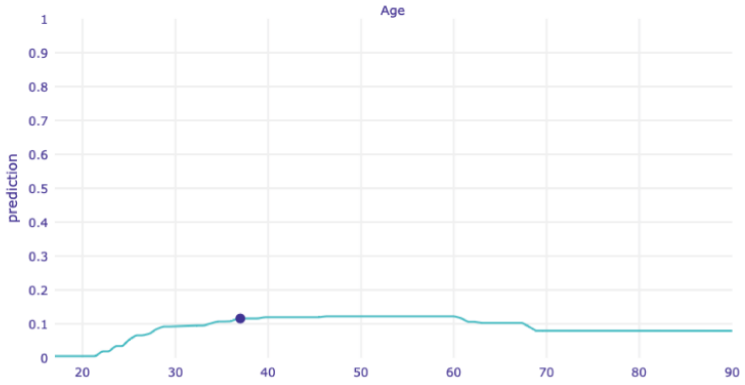
Woman 1

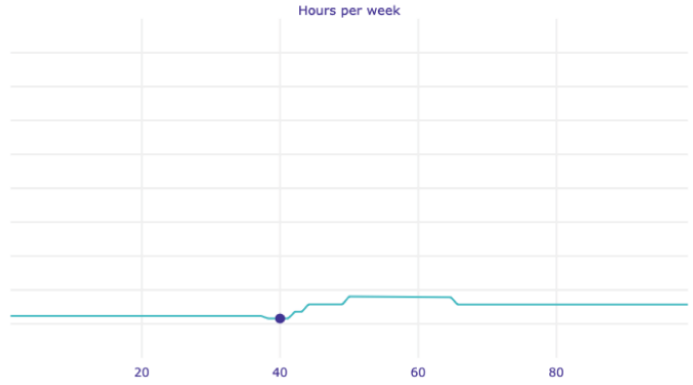
Below you will find the information of Applicant X.

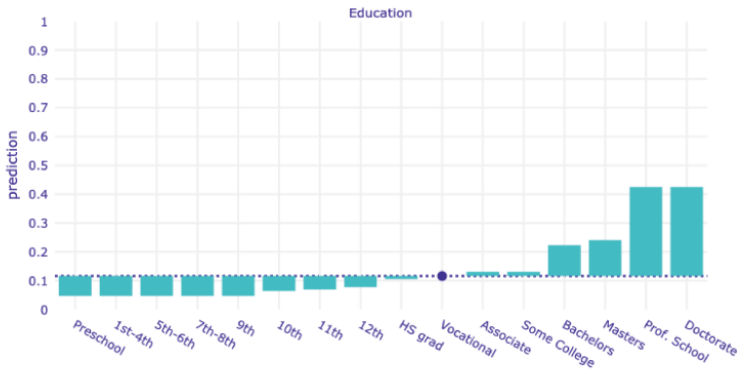
You can see that the model made a prediction of whether to approve or deny a loan from this applicant based on five factors. The explanation is below.

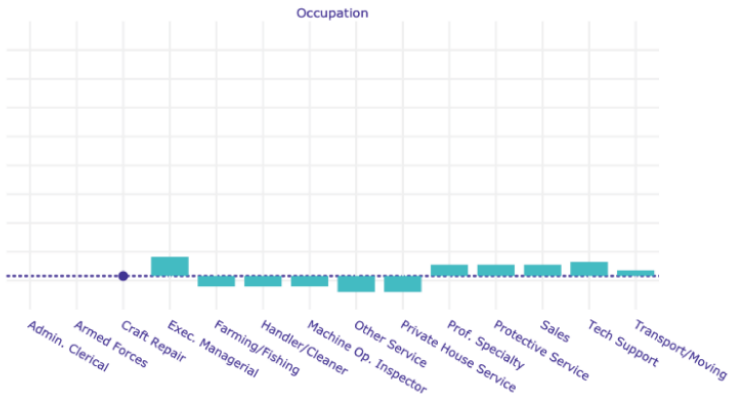
Look at the explanation, and answer the questions that follow.

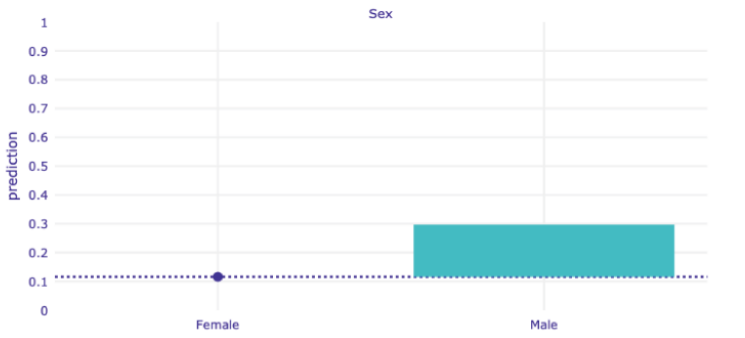
Remember that if the model's prediction probability (Predicted Value) for 'YES' is greater than or equal to 0.5, the model will return 'YES' (approve the loan). If it is less than 0.5, the model will return 'NO' (deny the loan).











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

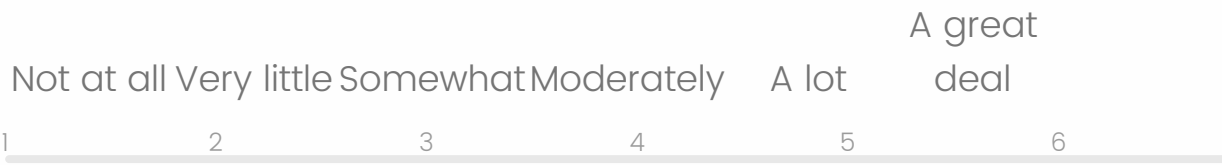
Which factor(s) are pushing the model toward predicting 'NO'?

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

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

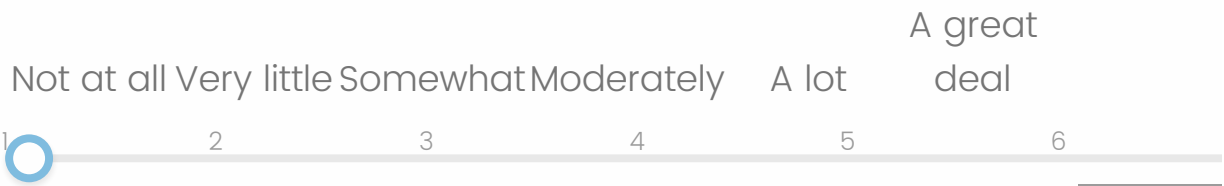
- [ ] Education
- [ ] Hours Worked Per Week
- [ ] Age
- [ ] Sex
- [ ] Occupation
- [ ] None of these

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



Level of Trust

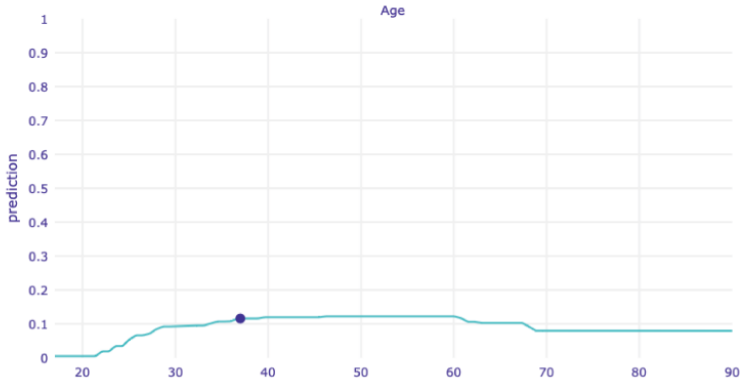
On a scale from 1 to 6, how much do you trust the model to approve or deny a loan for **other people in general**?

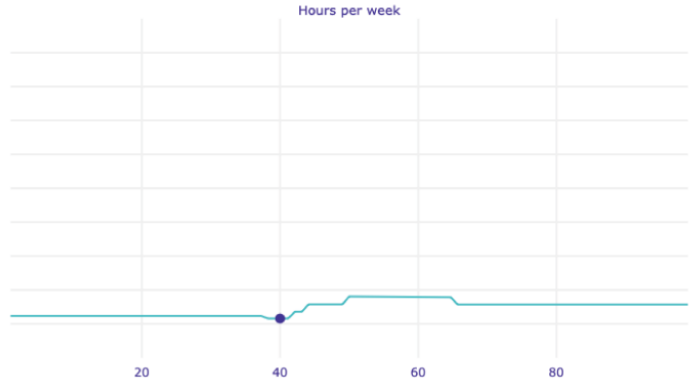


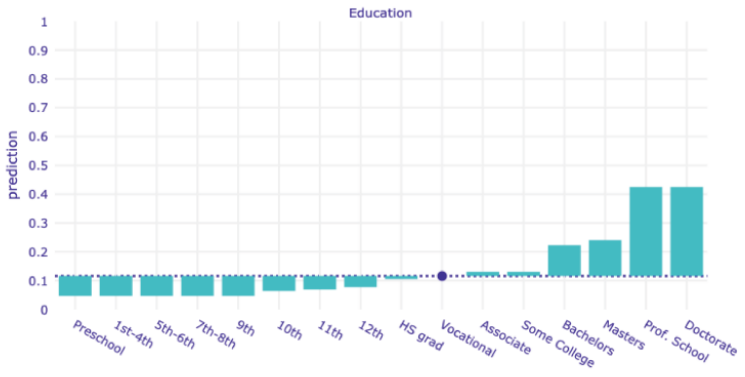
Level of Trust

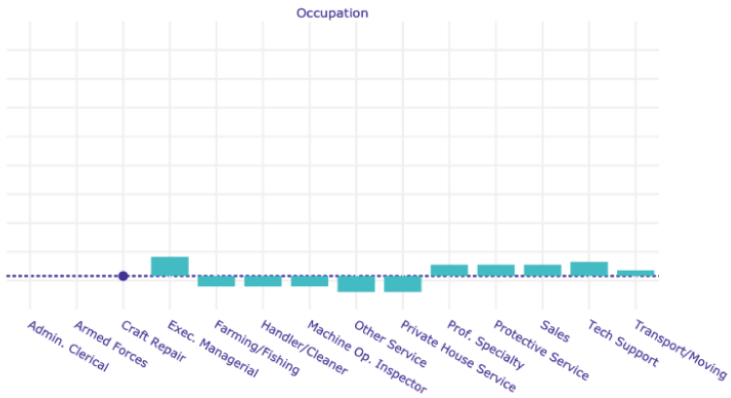
Please indicate whether you agree with the below statements.

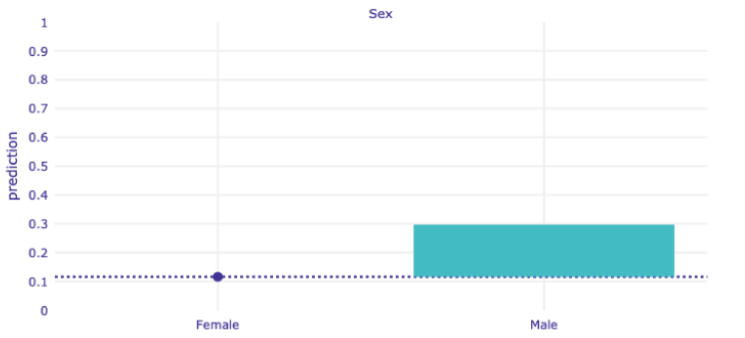
	Agree
This model uses all of the features that it should use when making this decision.	○
This model does not use any unnecessary features when making this decision.	○
I trust the data this model was trained on.	○
Computer models can be trusted to make human decisions.	○
This model is accurate.	○
This 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

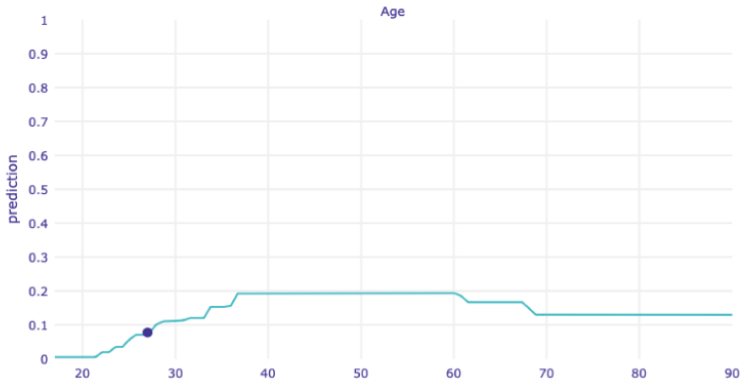
Below you will find the information of Applicant R.

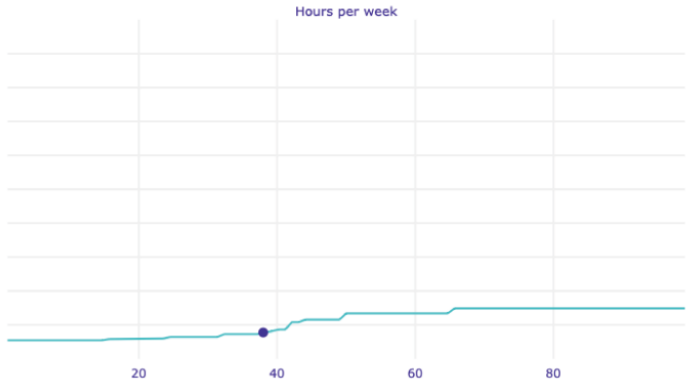
You can see that the model made a prediction of whether to approve or deny a loan from this applicant based on five factors. The explanation is below.

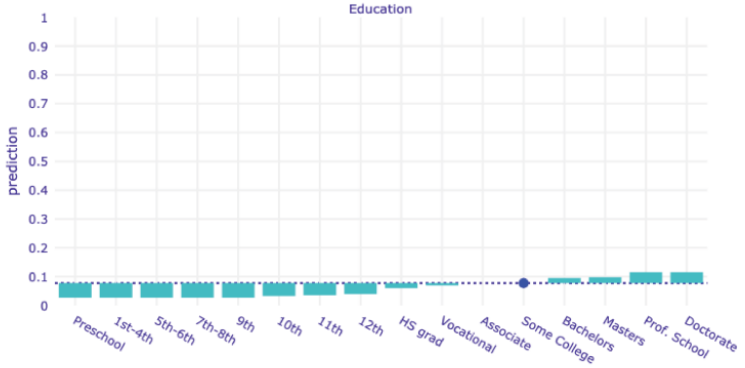
Look at the explanation, and answer the questions that follow.

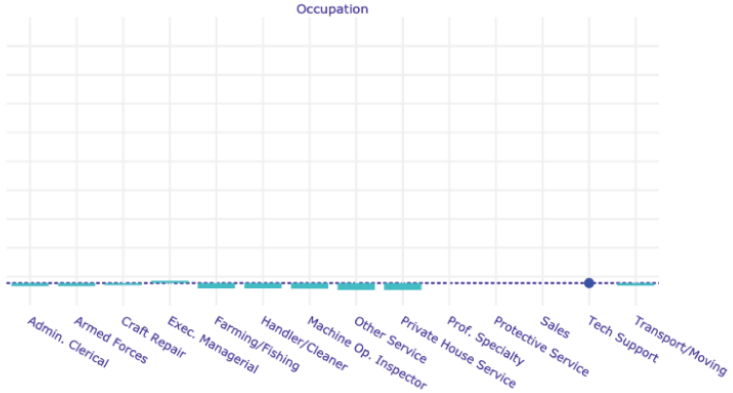
Remember that if the model's prediction probability (Predicted Value) for 'YES' is greater than or equal to 0.5, the model will return 'YES' (approve the loan). If it is less than 0.5, the model will

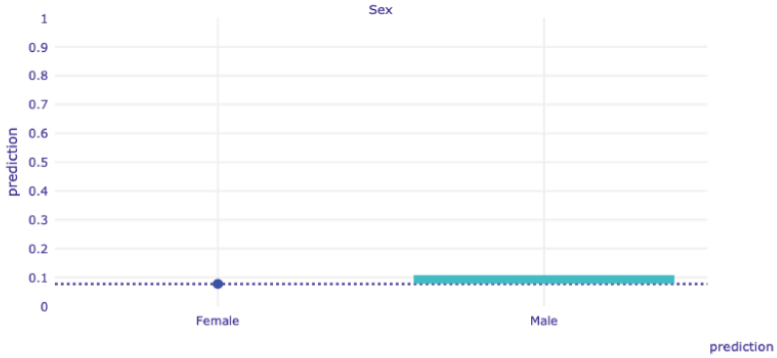
return 'NO' (deny the loan).











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**?

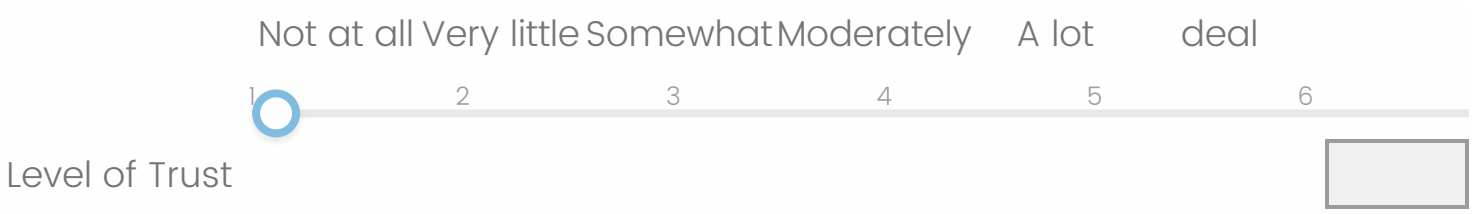
Not at all Very little Somewhat Moderately A lot A great deal

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**?

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.	○
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.	○

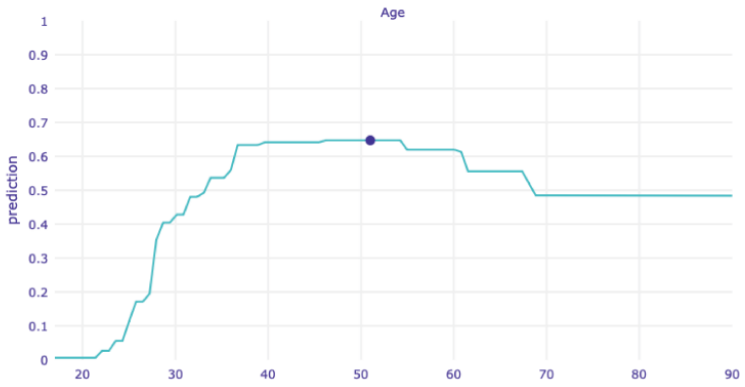
Woman 3

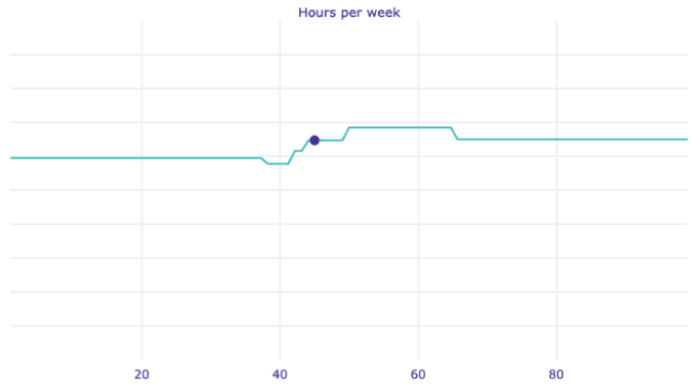
Below you will find the information of Applicant S.

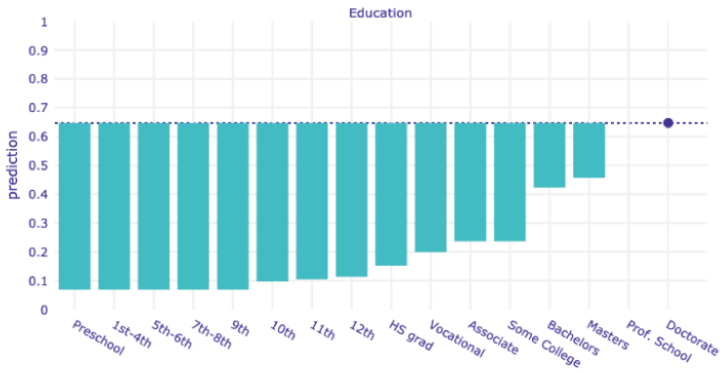
You can see that the model made a prediction of whether to approve or deny a loan from this applicant based on five factors. The explanation is below.

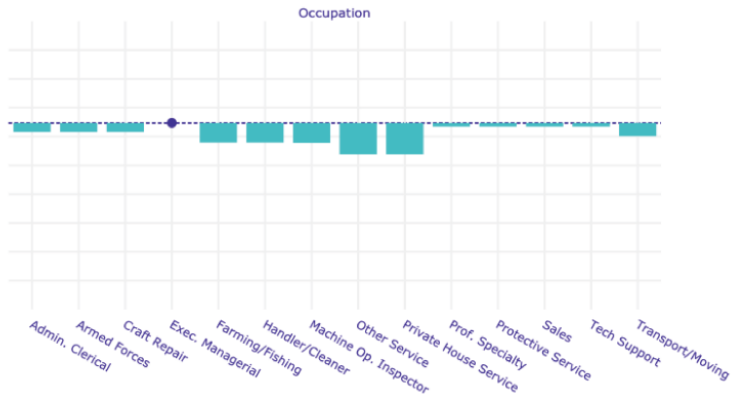
Look at the explanation, and answer the questions that follow.

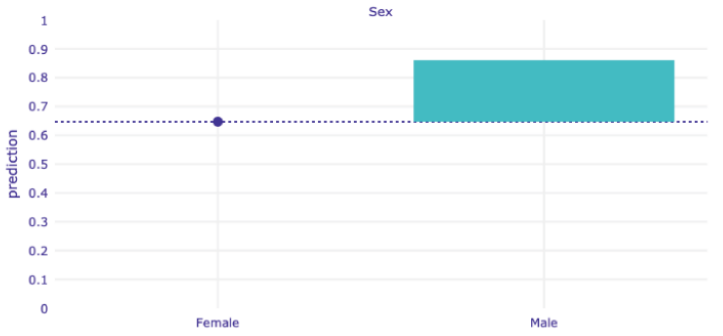
Remember that if the model's prediction probability (Predicted Value) for 'YES' is greater than or equal to 0.5, the model will return 'YES' (approve the loan). If it is less than 0.5, the model will return 'NO' (deny the loan).











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'?

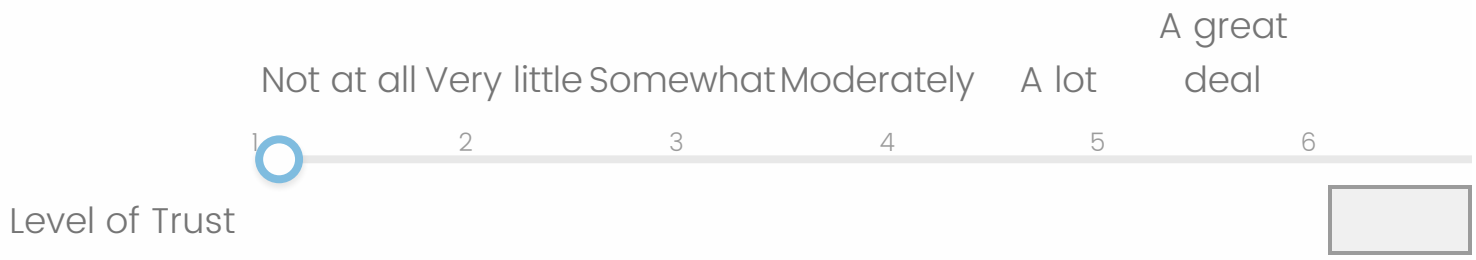
- ☐ Education
- ☐ Hours Worked Per Week
- ☐ Age
- ☐ Sex
- ☐ Occupation
- ☐ None of these

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

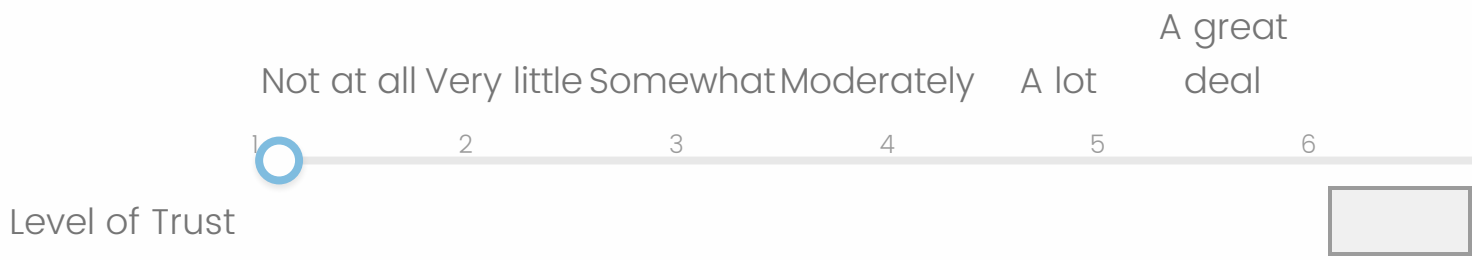
- ☐ Education

- [ ] 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**?



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 1

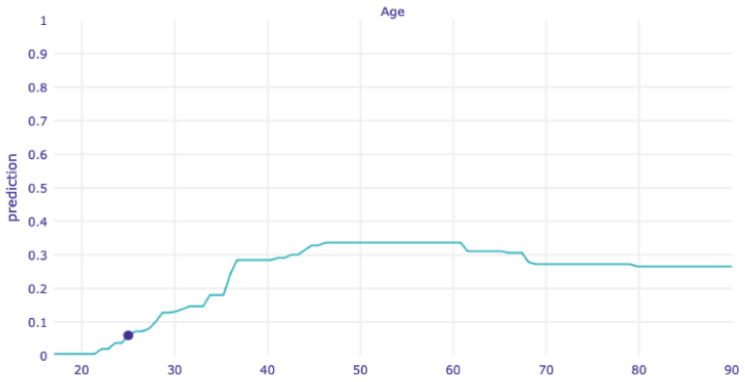
Below you will find the information of Applicant N.

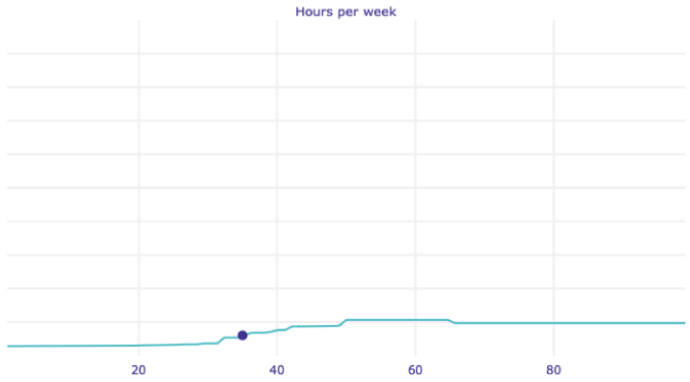
You can see that the model made a prediction of whether to

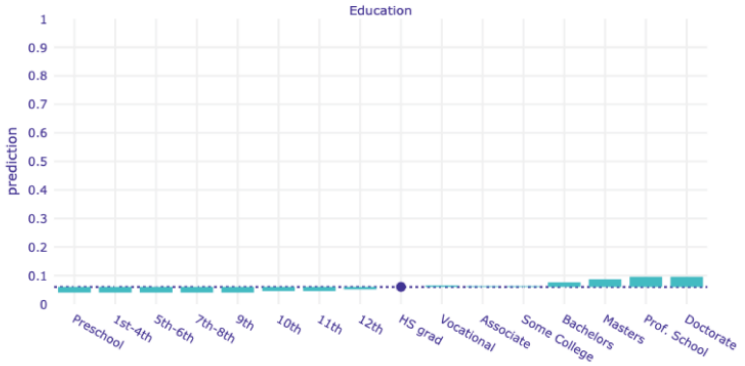
approve or deny a loan from this applicant based on five factors. The explanation is below.

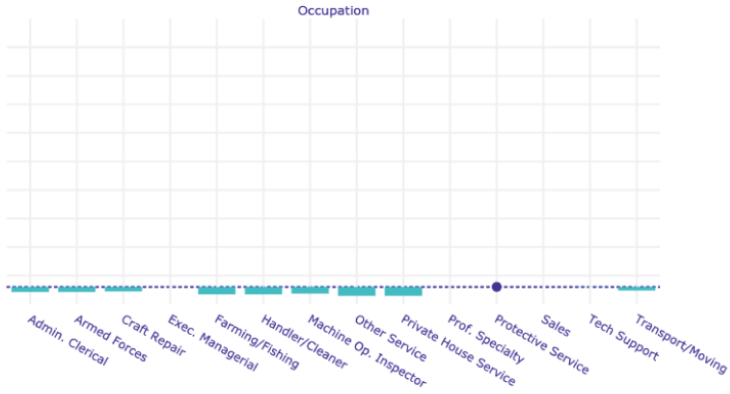
Look at the explanation, and answer the questions that follow.

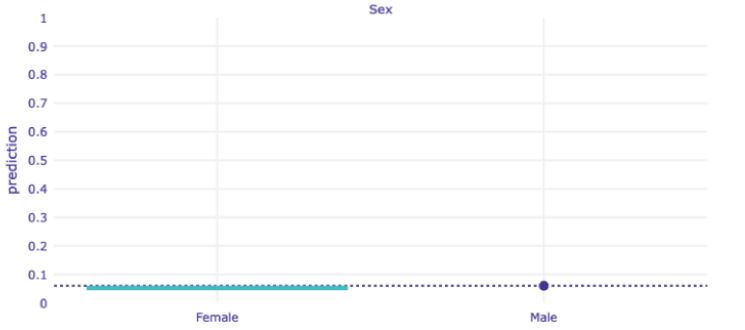
Remember that if the model's prediction probability (Predicted Value) for 'YES' is greater than or equal to 0.5, the model will return 'YES' (approve the loan). If it is less than 0.5, the model will return 'NO' (deny the loan).











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'?

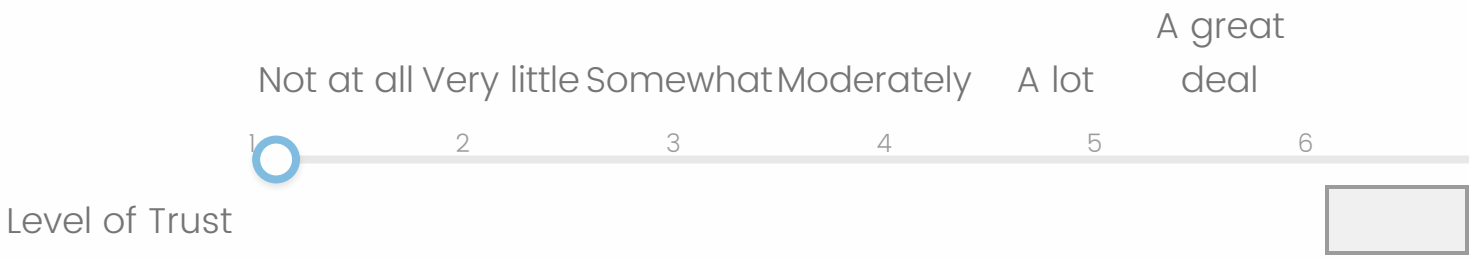
- ☐ Education
- ☐ Hours Worked Per Week
- ☐ Age
- ☐ Sex
- ☐ Occupation
- ☐ None of these

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

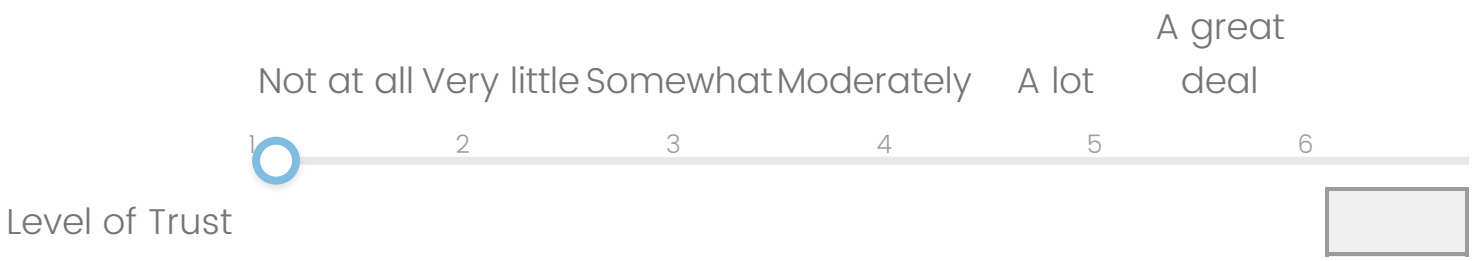
- ☐ Education
- ☐ Hours Worked Per Week

- [ ] 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**?



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.	○

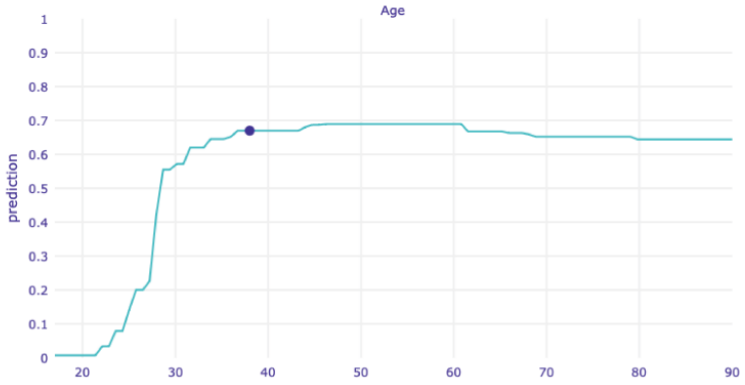
Man 2

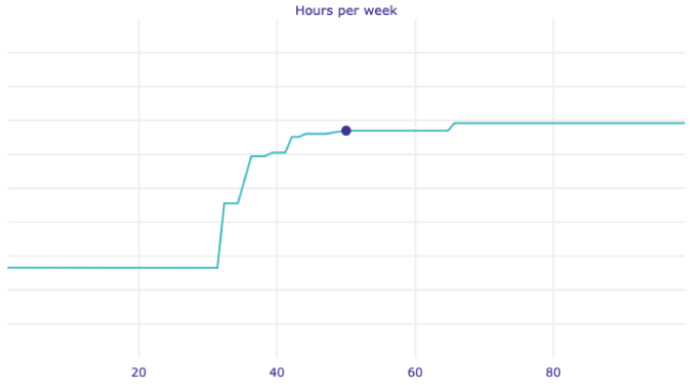
Below you will find the information of Applicant P.

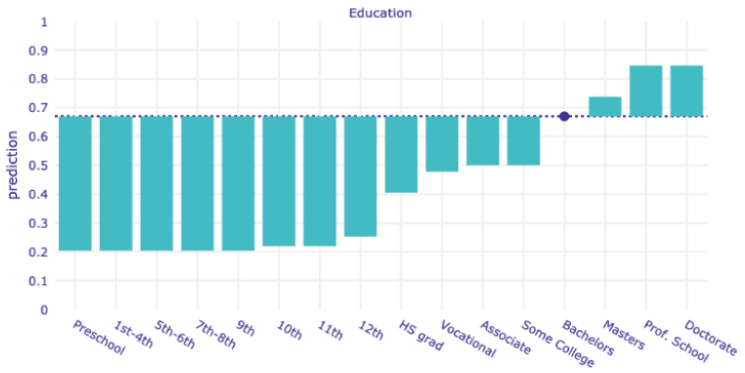
You can see that the model made a prediction of whether to approve or deny a loan from this applicant based on five factors.

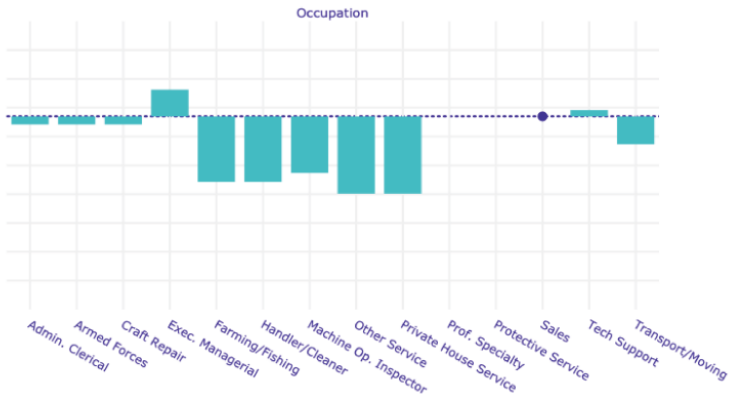
The explanation is below. Look at the explanation, and answer the questions that follow.

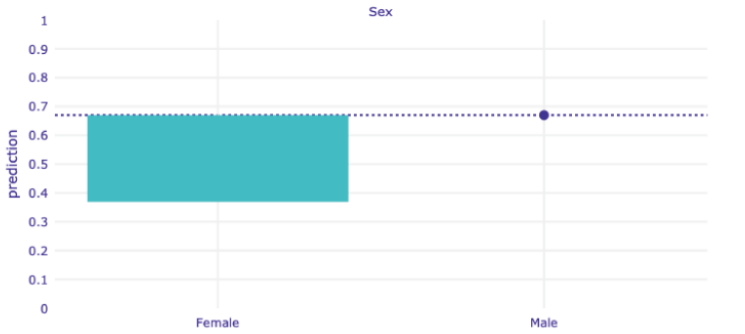
Remember that if the model's prediction probability (Predicted Value) for 'YES' is greater than or equal to 0.5, the model will return 'YES' (approve the loan). If it is less than 0.5, the model will return 'NO' (deny the loan).











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'?

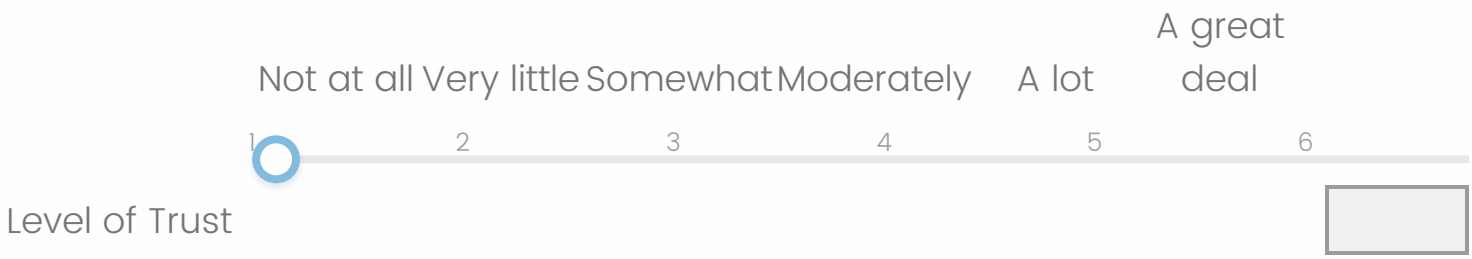
- ☐ Education
- ☐ Hours Worked Per Week
- ☐ Age
- ☐ Sex
- ☐ Occupation
- ☐ None of these

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

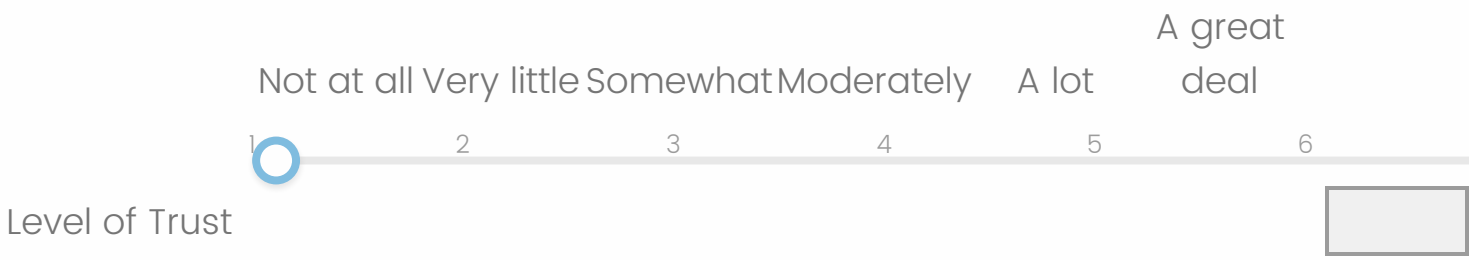
- ☐ Education

- [ ] 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

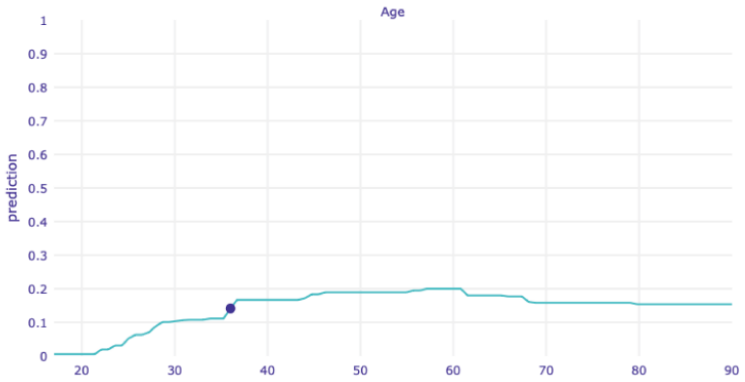
Below you will find the information of Applicant K.

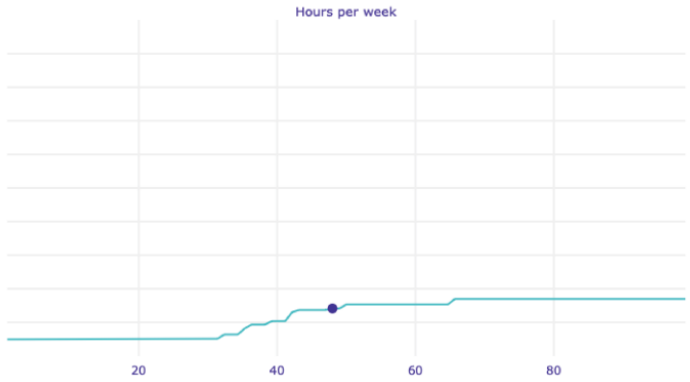
You can see that the model made a prediction of whether to

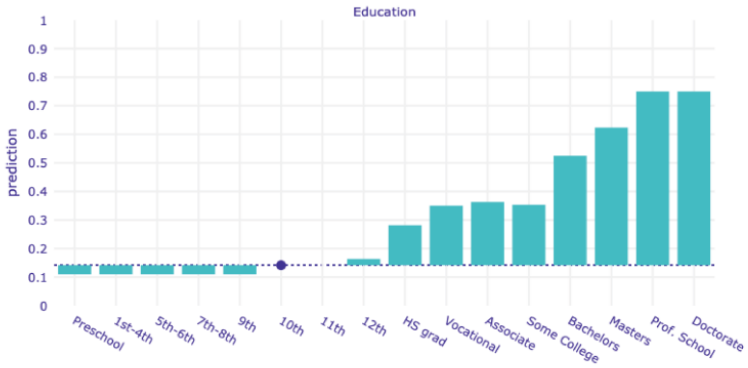
approve or deny a loan from this applicant based on five factors. The explanation is below.

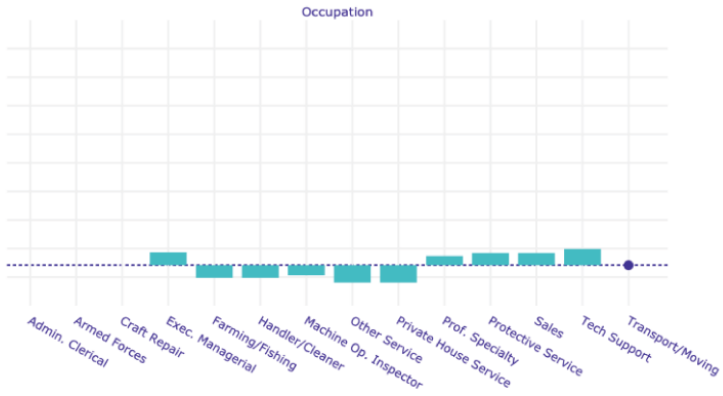
Look at the explanation, and answer the questions that follow.

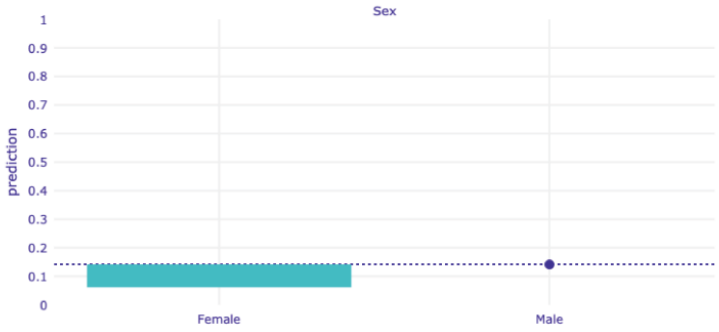
Remember that if the model's prediction probability (Predicted Value) for 'YES' is greater than or equal to 0.5, the model will return 'YES' (approve the loan). If it is less than 0.5, the model will return 'NO' (deny the loan).











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.

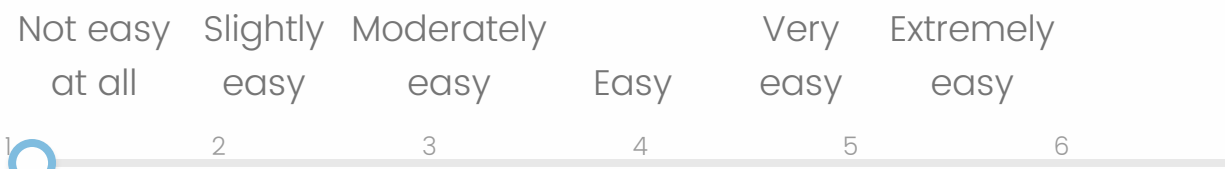
	Agree
This model uses all of the features that it should use when making this decision.	○
This model does not use any unnecessary features when making this decision.	○
I trust the data this model was trained on.	○
Computer models can be trusted to make human decisions.	○
This model is accurate.	○
This 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

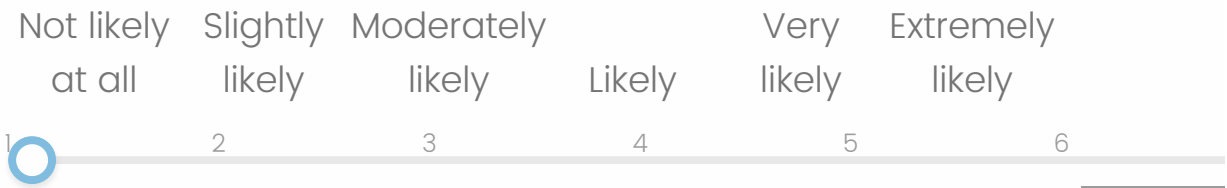
How well did you understand the way this model makes decisions?

Not well at all	Slightly well	Moderately well	Well	Very well	Extremely well

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



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

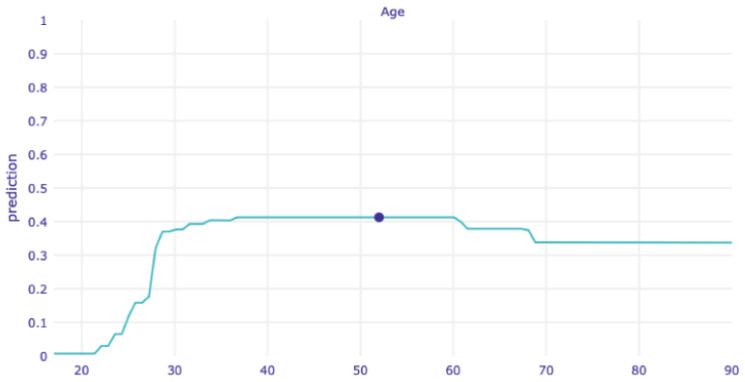


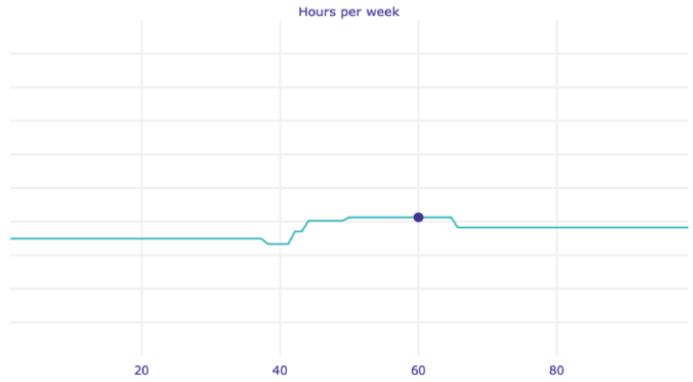
Fairness

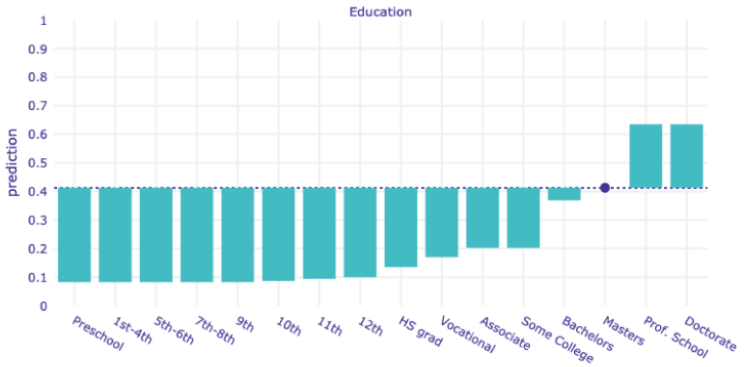
Below are two explanations for predictions made by the same loan approval machine learning model you have been seeing, for two people with almost identical features.

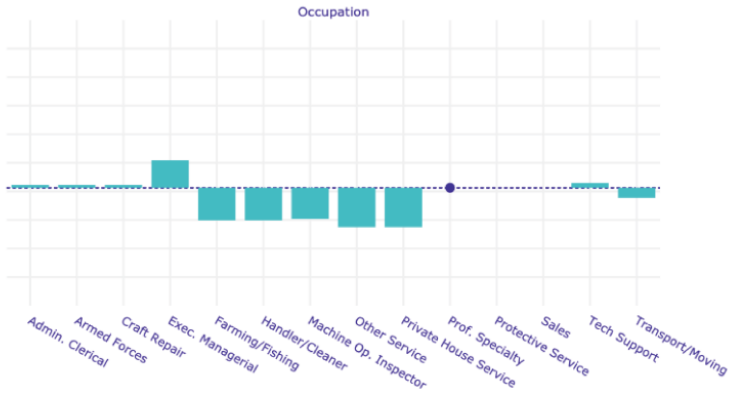
Remember that if the model's prediction probability (Predicted Value) for 'YES' is greater than or equal to 0.5, the model will return 'YES' (approve the loan). If it is less than 0.5, the model will return 'NO' (deny the loan).

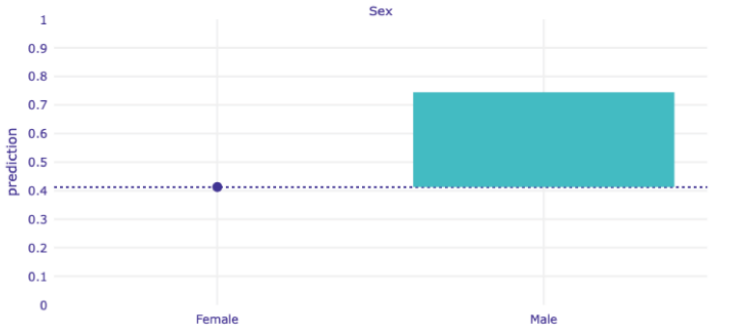
Person A


Age
prediction
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
20
30
40
50
60
70
80
90

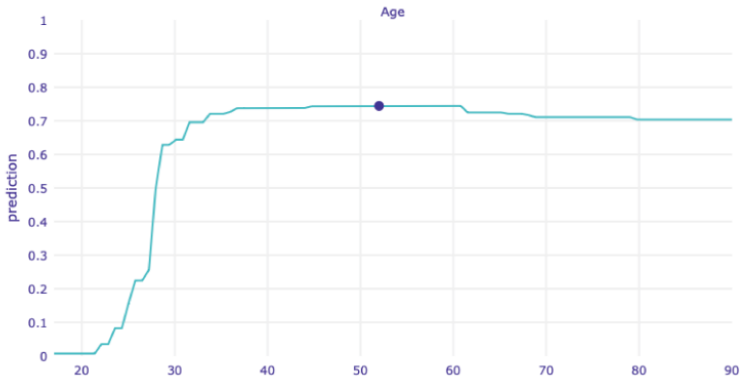

Hours per week
20
40
60
80

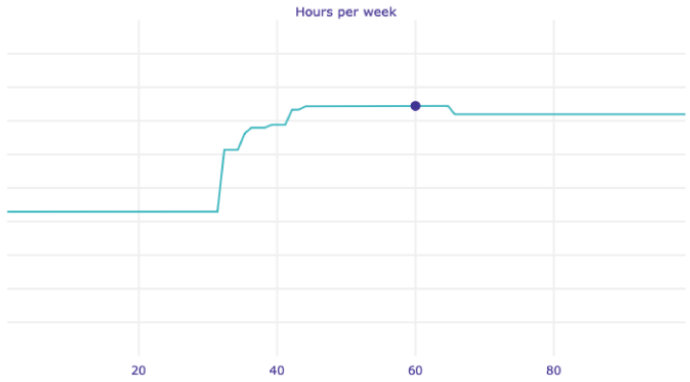

Education
prediction
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Preschool
1st-4th
5th-6th
7th-8th
9th
10th
11th
12th
HS grad
Vocational
Associate
Some College
Bachelors
Masters
Prof. School
Doctorate

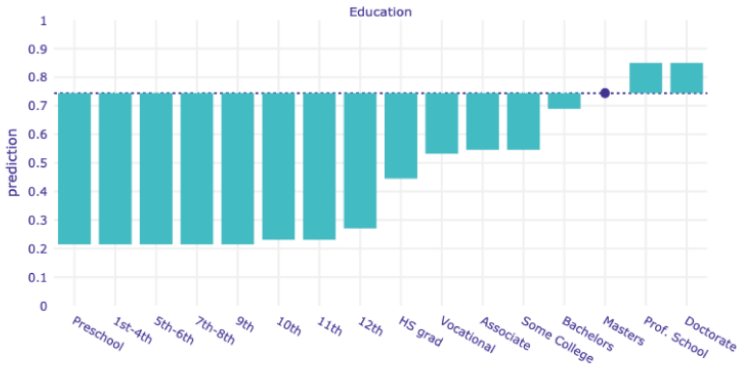

Occupation
Admin. Clerical
Armed Forces
Craft Repair
Exec. Managerial
Farming/Fishing
Handler/Cleaner
Machine Op. Inspector
Other Service
Private House Service
Prof. Specialty
Protective Service
Sales
Tech Support
Transport/Moving

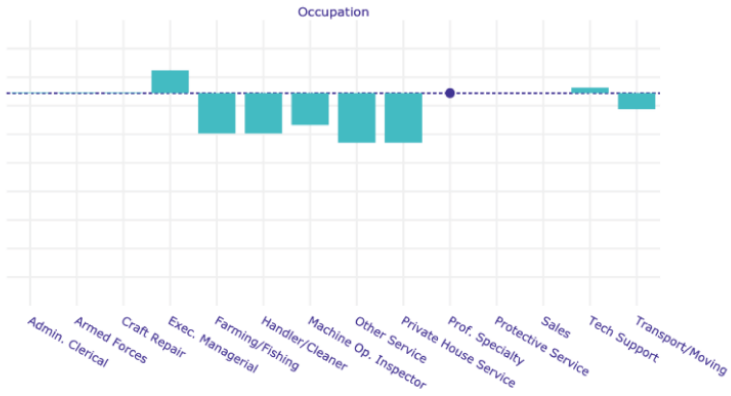

Sex
prediction
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Female
Male

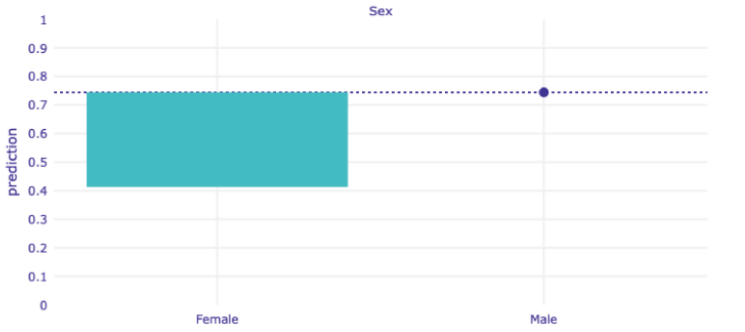
Person B











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**?

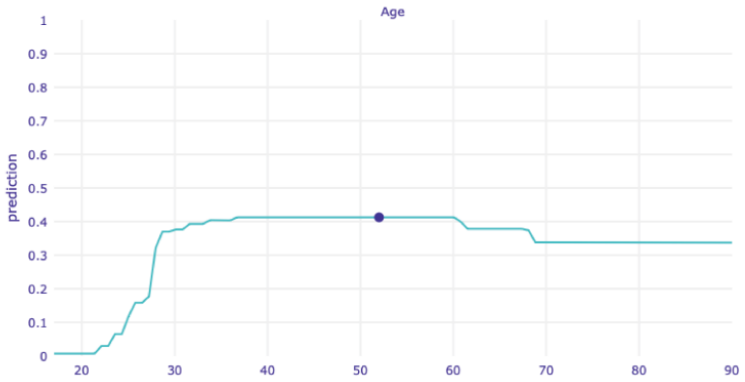
	Not at all	Very little	Somewhat	Moderately	A lot	A great deal
	1	2	3	4	5	6
Level of Trust						

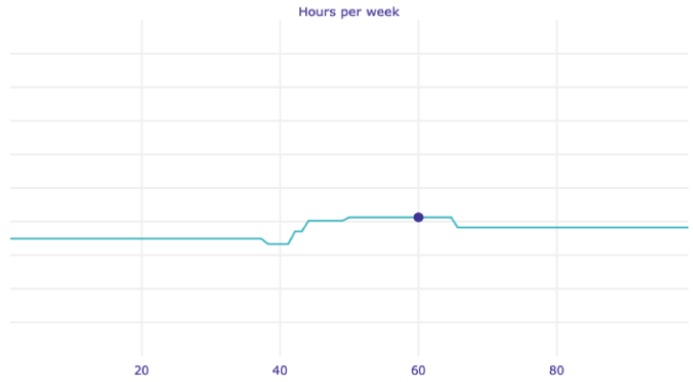
Please indicate whether you agree with the below statements.

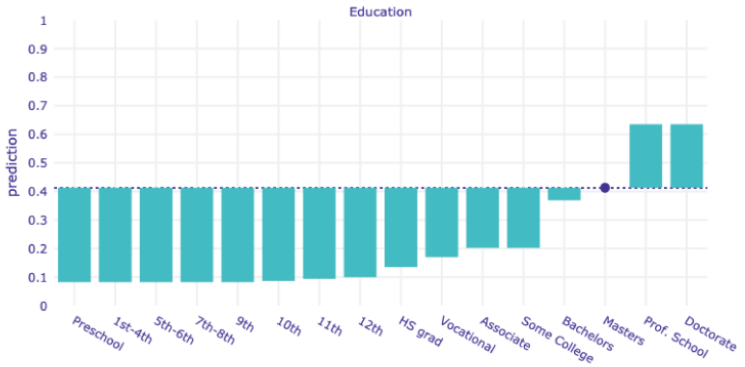
	Agree
This model uses all of the features that it should use when making this decision.	○
This model does not use any unnecessary features when making this decision.	○
I trust the data this model was trained on.	○
Computer models can be trusted to make human decisions.	○
This model is accurate.	○
This 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.	○

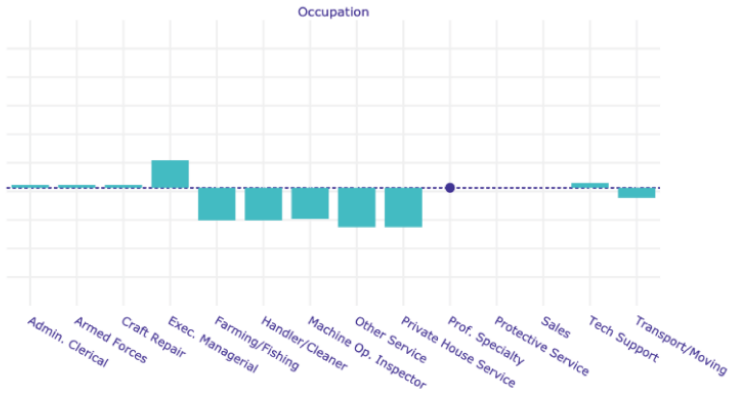
Fairness General

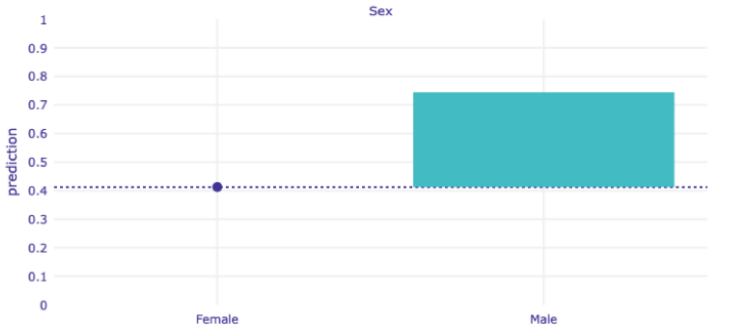
Person A


Age
prediction
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
20
30
40
50
60
70
80
90

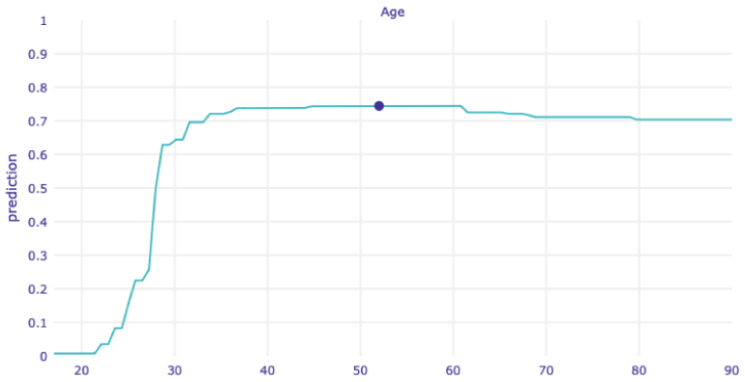

Hours per week
20
40
60
80

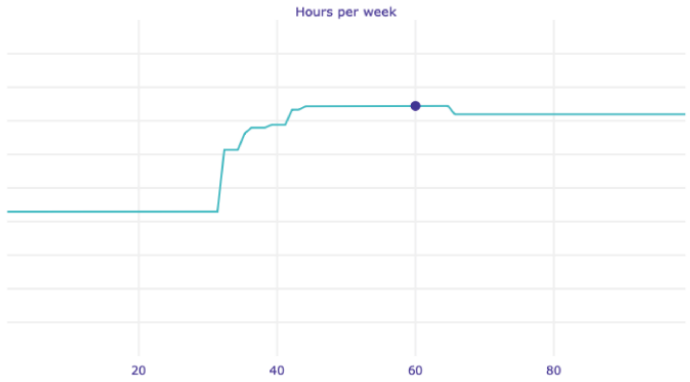

Education
prediction
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Preschool
1st-4th
5th-6th
7th-8th
9th
10th
11th
12th
HS grad
Vocational
Associate
Some College
Bachelors
Masters
Prof. School
Doctorate

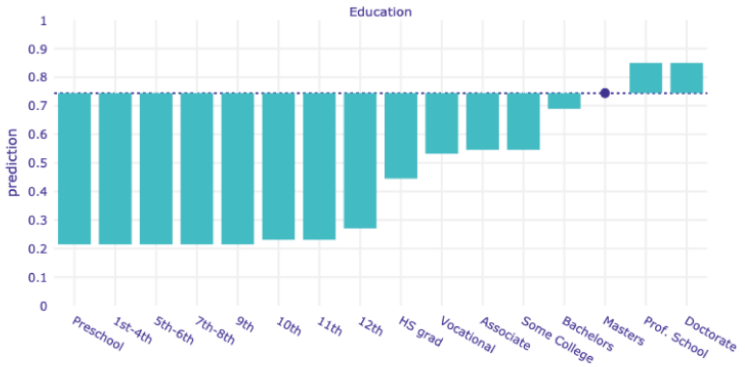

Occupation
Admin. Clerical
Armed Forces
Craft Repair
Exec. Managerial
Farming/Fishing
Handler/Cleaner
Machine Op. Inspector
Other Service
Private House Service
Prof. Specialty
Protective Service
Sales
Tech Support
Transport/Moving

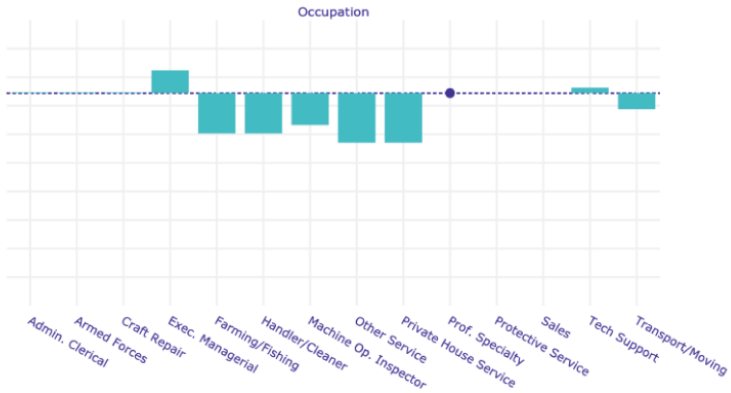

Sex
prediction
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Female
Male

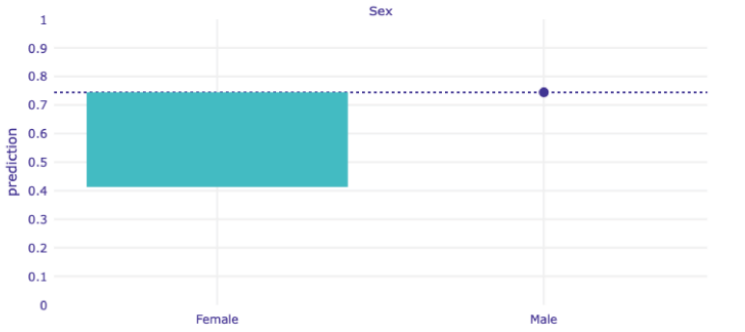
Person B











Do you think this model includes potentially discriminating factors?

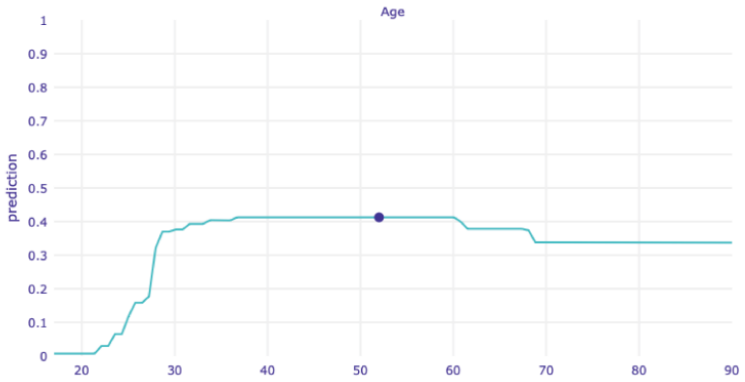
- ○ YES

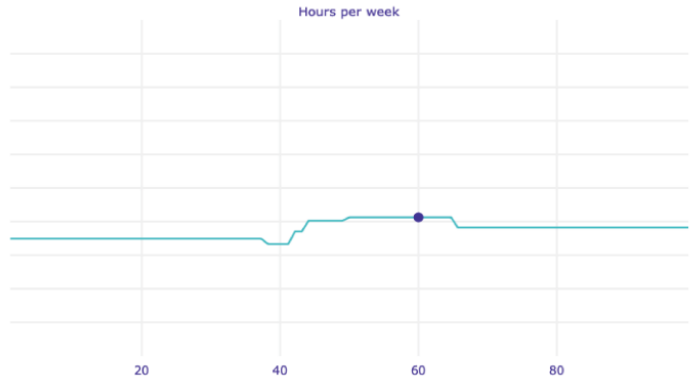
○ NO

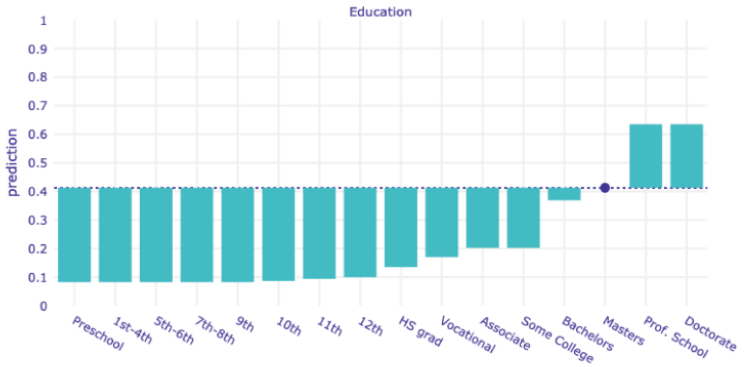
If yes, which ones?

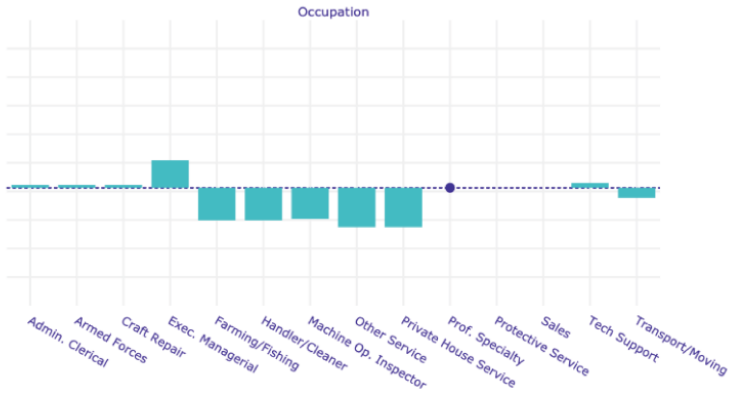
- ☐ Age
- ☐ Hours Per Week
- ☐ Education
- ☐ Occupation
- ☐ Sex

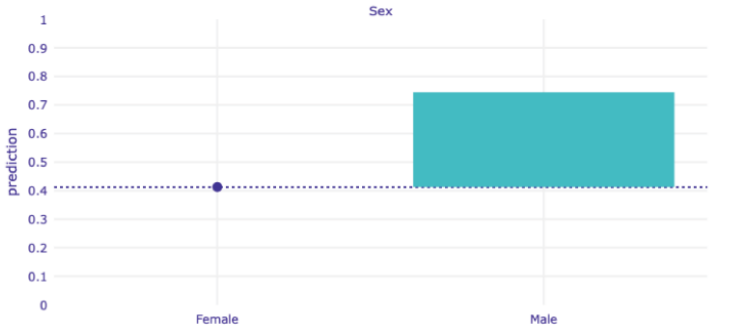
Person A


Age
prediction
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
20
30
40
50
60
70
80
90

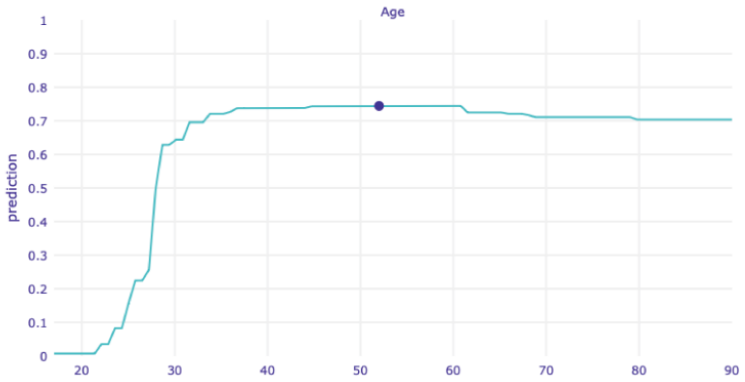

Hours per week
20
40
60
80

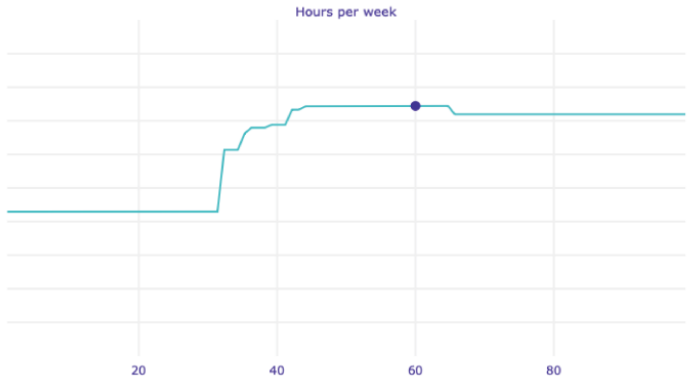

Education
prediction
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Preschool
1st-4th
5th-6th
7th-8th
9th
10th
11th
12th
HS grad
Vocational
Associate
Some College
Bachelors
Masters
Prof. School
Doctorate

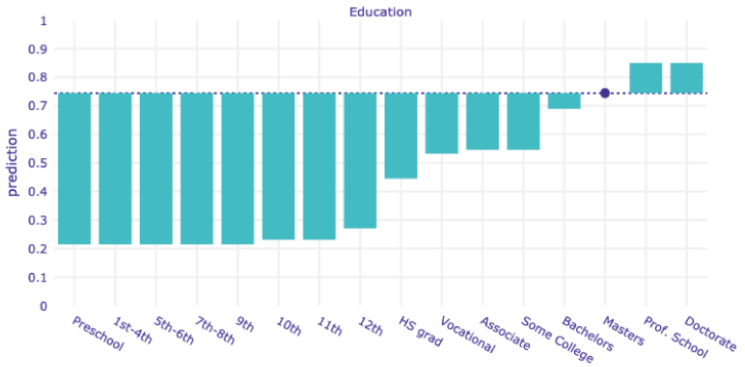

Occupation
Admin. Clerical
Armed Forces
Craft Repair
Exec. Managerial
Farming/Fishing
Handler/Cleaner
Machine Op. Inspector
Other Service
Private House Service
Prof. Specialty
Protective Service
Sales
Tech Support
Transport/Moving

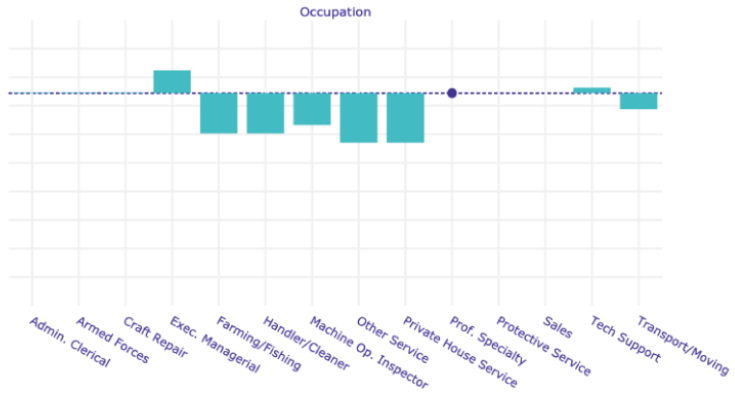

Sex
prediction
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Female
Male

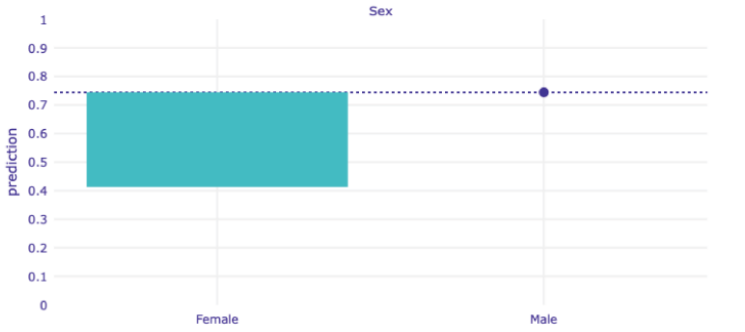
Person B











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