Value Cards

Model Cards

8 Models
1 Model Overview

Persona Cards

Judge
Defendants
Fairness Advocate
Community Member

Checklist Card

Understanding Societal Values in AI
Identifying Stakeholders
Analyzing Impacts

Model Number: 1

Accuracy

Over all: 68.4%

White American: 71.3%

African American: 65.5%

Disparity

Disparity in Accuracy: 5.8%

Disparity in FPR: 6.1%

Disparity in FNR: 14.9%

False Positive Rate

Over all: 8.0%

White American: 5.2%

African American: 11.3%

False Negative Rate

Over all: 74.6%

White American: 83.2%

African American: 68.3%

Model Number: 2

Accuracy

Over all: 60.0%

White American: 60.0%

African American: 60.0%

Disparity

Disparity in Accuracy: 0%

Disparity in FPR: 6.4%

Disparity in FNR: 9.2%

False Positive Rate

Over all: 45.5%

White American: 42.1%

African American: 49.5%

False Negative Rate

Over all: 29.8%

White American: 35.1%

African American: 25.9%

Model Number: 3

Accuracy

Over all: 70.4%

White American: 72.9%

African American: 68.0%

Disparity

Disparity in Accuracy: 4.9%

Disparity in FPR: 7%

Disparity in FNR: 9%

False Positive Rate

Over all: 10.1%

White American: 5.7%

African American: 15.1%

False Negative Rate

Over all: 65.2%

White American: 77.1%

African American: 56.5%

Model Number: 4

Accuracy

Over all: 64.1%

White American: 65.2%

African American: 63.1%

Disparity

Disparity in Accuracy: 2.1%

Disparity in FPR: 6.7%

Disparity in FNR: 9.8%

False Positive Rate

Over all: 31.9%

White American:28.9%

African American: 35.6%

False Negative Rate

Over all: 42.8%

White American: 48.5%

African American: 38.7%

Model Number: 5

Accuracy

Over all: 66.7%

White American: 70.6%

African American: 62.7%

Disparity

Disparity in Accuracy: 7.9%

Disparity in FPR: 2.6%

Disparity in FNR: 9.8%

False Positive Rate

Over all: 1.9%

White American: 0.8%

African American: 3.4%

False Negative Rate

Over all: 90.5%

White American: 96.2%

African American: 86.4%

Model Number: 6

Accuracy

Over all: 67.6%

White American: 69.9%

African American: 65.2%

Disparity

Disparity in Accuracy: 4.7%

Disparity in FPR: 18.0%

Disparity in FNR: 23.2%

False Positive Rate

Over all: 30.6%

White American: 21.9%

African American: 40.8%

False Negative Rate

Over all: 35.6%

White American: 49.0%

African American: 25.8%

Model Number: 7

Accuracy

Over all: 69.1%

White American: 72.6%

African American: 65.4%

Disparity

Disparity in Accuracy: 7.2%

Disparity in FPR: 6.9%

Disparity in FNR: 15.1%

False Positive Rate

Over all: 4.7%

White American: 1.5%

African American: 8.4%

False Negative Rate

Over all: 78.7%

White American: 87.4%

African American: 72.3%

Model Number: 8

Accuracy

Over all: 60%

White American: 60.8%

African American: 57.3%

Disparity

Disparity in Accuracy: 3.5%

Disparity in FPR: 19.4%

Disparity in FNR: 14.6%

False Positive Rate

Over all: 53.5%

White American: 44.6%

African American: 64%

False Negative Rate

Over all: 18%

White American: 26.6%

African American: 12%

Attribute / Model	Accuracy	FPR	FNR	Dis/Acc	Dis/FPR	Dis/FNR
1	☆					
2	✖	✖	☆	☆		
3	☆					
4				☆		
5		☆	✖		☆	
6					✖	✖
7	☆	☆	✖			
8	✖	✖	☆	☆	✖	

Accuracy Range:
43% - 71%

FNR Range:
3.4% - 94%

FPR Range:
1% - 87%

✖ performs relatively badly

☆ performs relatively well

Disclaimer: The performances are evaluated among the models, and only serve as a subjective reference. Please **always** refer to the model cards for model details.

Persona: Community Member

As a community member, you want your neighborhood to be safe. You are concerned about re-offending behavior in your community. In this case, you may care mostly about **false negative rate**.

However, your teammates, who play other roles in the system, might prioritize other metrics. Throughout the deliberation, please express your value, respect their concerns, and negotiate with them.

Persona: Fairness Advocate

As a fairness advocate, you want to prevent the unfair treatment on some demographic. You don't want the recidivism algorithm to be biased against one demographic. In this case, you want the model to have a **lower disparity**.

However, your teammates, who play other roles in the system, might prioritize other metrics. Throughout the deliberation, please express your value, respect their concerns, and negotiate with them.

Persona: Defendants

As a defendant, you are most worried about being falsely predicted as “will offend again.” In this case, you probably want the model to have a **lower false positive rate.**

However, your teammates, who play other roles in the system, might prioritize other metrics. Throughout the deliberation, please express your value, respect their concerns, and negotiate with them.

Persona: Judge

As a judge, you care most about making the right decision when sentencing a defendant. You probably want the model to have a **higher accuracy.**

However, your teammates, who play other roles in the system, might prioritize other metrics. Throughout the deliberation, please express your value, respect their concerns, and negotiate with them.

Checklists

1. Understanding societal values in AI

- There is no single definition of societal values that will apply equally well to different applications of AI.
- Prioritizing one value in AI systems often means making tradeoffs on competing values. It is therefore important to be explicit and transparent about priorities and assumptions.
- There are seldom clear-cut answers. It is therefore important to document your processes and considerations (including priorities and tradeoffs).

Checklists

2. Identifying stakeholders

- **Who is at risk of experiencing impacts**: considering both the people who will use the system and the people who will be directly or indirectly affected by the system, either by choice or not.
- **People often belong to overlapping groups**—different combinations of race, gender, and age, for example—and specific intersectional groups may be at greatest risk of experiencing different types of harm.
- **What are the tradeoffs** between expected benefits and potential harms for identified stakeholder groups?

Checklists

3. Analyzing impacts

- What are the **types of impact** (e.g., allocation, quality of service, stereotyping, denigration, over- or underrepresentation) on different stakeholders
- Rate the **degree** of impact [no discernable - minor - moderate - major]
- Estimate the **scale** of impact [small - medium - large]
- Estimate the **overall direction** of impact [positive - mostly positive - mostly negative - negative]

Quiz (Sept 10th)

⚠ This is a preview of the published version of the quiz

Started: Dec 14 at 3:15pm

Quiz Instructions

In this quiz, we will focus on an algorithm that makes **predictions about the likelihood of a criminal defendant's *recidivism***. Recidivism refers to a criminal defendant who eventually commits another crime. The algorithm predicts whether a defendant will relapse into criminal behavior, and this prediction can be used by a judge in determining the defendant's sentence.

The algorithm learns from **historical criminal recidivism data** to predict the likelihood that a defendant will reoffend. The basic idea is that the algorithm **considers a defendant likely to reoffend if his/her profile is similar to the profiles of other defendants who have reoffended.**

The algorithm uses a number of attributes of defendants, such as their age, gender, previous criminal records, etc, to evaluate defendants' recidivism risk. The algorithm may weigh some attributes more heavily than others in making its prediction.

Let us remind you of the definitions of performance metrics:

True positive (TP): people who are being predicted to re-offend and actually re-offend

True negative (TN): people who are being predicted to NOT reoffend and actually do NOT re-offend

False positive (FP): people who are being predicted to re-offend but actually do NOT re-offend

False negative (FN): people who are being predicted to NOT re-offend but actually re-offend

True Positive Rate (TPR): The fraction of people who are predicted correctly to re-offend, among all those who actually re-offend. (TP / TP + FN)

True Negative Rate (TNR): The fraction of people who are predicted correctly to NOT reoffend, among all those who actually do NOT re-offend. (TN/ TN + FP)

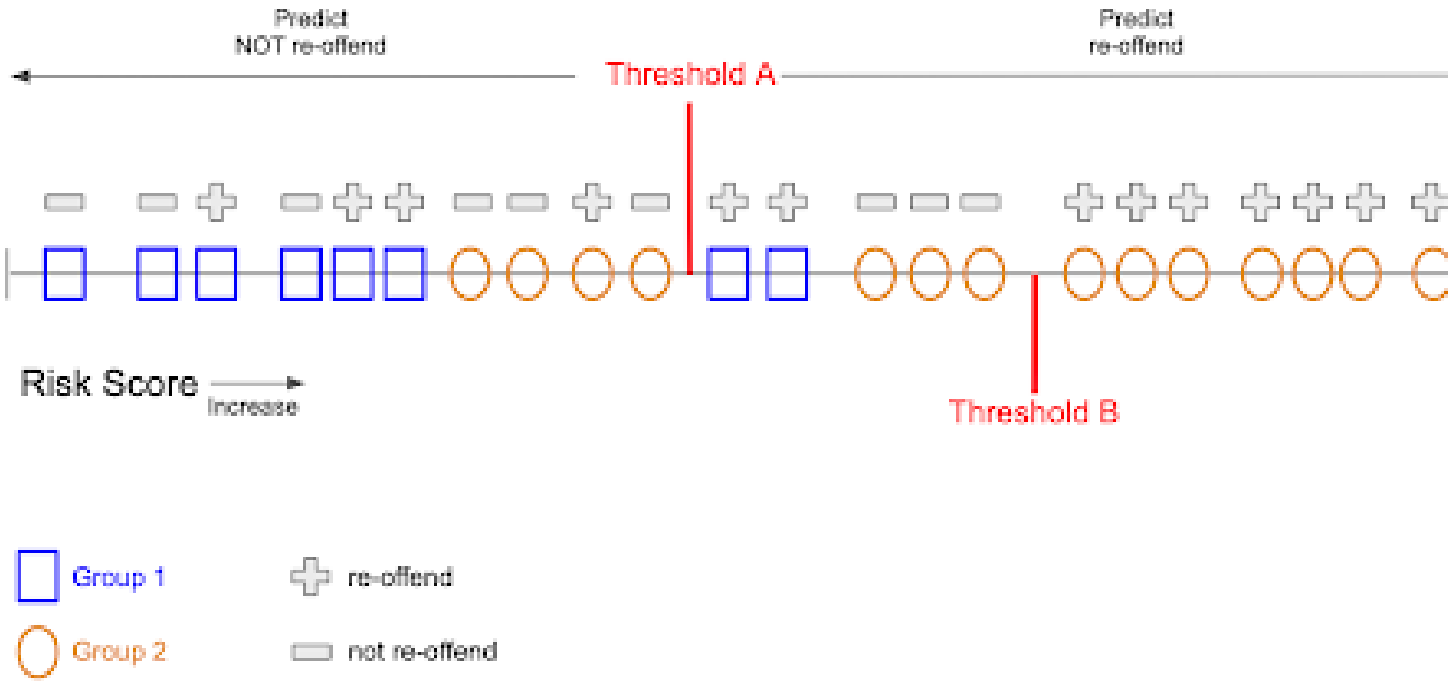
False Positive Rate (TPR): The fraction of people who are falsely predicted to re-offend, among all those who do NOT re-offend (FP/FP+ TN).

False Negative Rate (FNR): The fraction of people who are falsely predicted to NOT re-offend, among all those who do re-offend (FN/FN+TP).

Accuracy: The percentage of defendants that are correctly predicted
(Correct Predictions) / (Total Predictions) = (TP+TN)/ (TP+TN + FP + FN)

Disparity: The difference in accuracy/error rates (e.g., false positive rates and false negative rates) between the two groups

Background



In this quiz, we will consider a specific type of prediction method that first provides a risk score for each individual defendant and then makes predictions by taking a threshold on the risk scores. As shown in the figure above, there are two groups (blue square and orange circle). Each individual defendant has a label (+ or -) that indicates whether such an individual reoffends or not. After choosing a threshold, the model will predict “re-offend” for any individual with a risk score higher than the threshold and predict “not re-offend” for any individual with a risk score lower than the threshold.

We will look at the effects on different error rates by changing the threshold.

For sanity check, there are 8 squares and 16 circles in the above picture.

Question 1 — 1 pts

If we pick **threshold A**, calculate the following:

What is the number of false negatives (FN) for the whole dataset:

What is the number of the true negatives (TN) for the whole dataset:

What is the false-negative rate (FNR) for the whole dataset (expressed as a ratio) :

What is the false-negative rate (FNR) for the Group 1 (blue square) (expressed as a ratio):

Question 2

1 pts

Which of threshold **A** and threshold **B** is optimal in terms of accuracy?

- ◯ Neither of them are optimal.
- ◯ Threshold A is optimal in terms of accuracy.
- ◯ Both of them are optimal.
- ◯ Threshold B is optimal in terms of accuracy.

Question 3

1 pts

Which of the following **reduces false positive rate**? [FPR=FP/N]?

- ◯ The model chooses a lower threshold.
- ◯ The model chooses a higher threshold.
- ◯ There is nothing you can do to minimize the false positive rate.

Question 4

1 pts

If your goal is to **reduce disparity between FPRs of the two groups**, what type of model would you pick?

- ◯ Pick the model where one demographic group has a higher false negative rate over another one.
- ◯ Pick the model that has similar false positive rates and false negative rates between two demographic groups.
- ◯ Pick the model where one demographic group has a higher false positive rate over another one.

Question 5

1 pts

Suppose you want to change the threshold to reduce false-negative rate. How might this affect the false-positive rate? (Check all that applies.)

- ☐ The false positive rate drops.
- ☐ The false positive rate increases.
- ☐ There will be no change in the false positives rate.

Question 6

1 pts

Suppose you want to change the threshold in the figure to minimize the disparity of false negative rates between two different demographic groups. How might this affect the algorithm's performance?

- ◯ The percentage of correct predictions will stay the same.
- ◯ The accuracy will be lower than the optimal accuracy.
- ◯ The accuracy will still be optimal.

Question 7

0 pts

For a recidivism prediction algorithm, what metrics below do you think are the most important ones to consider in tuning the algorithm? Assume there are tradeoffs between these metrics so that you can not optimize for all of them at the same time. [There are no correct answers in this one]

- ☐ Minimizing False Negative Rate
- ☐ Maximizing True Negative Rate
- ☐ Minimizing disparity (of FPRs, FNRs, accuracy rates) between demographics
- ☐ Minimizing False Positive Rate
- ☐ Maximizing Accuracy
- ☐ Maximizing True Positive Rate
- ☐ Other

Question 8

1 pts

Please explain your reasoning for the previous question.

HTML Editor

12pt Paragraph

0 words

Question 9

0 pts

In any algorithmic decision-making system, what metrics below do you think are the most important ones to consider in tuning the algorithm? [There are no correct answers in this one]

- [ ] Maximizing True Negative Rate
- [ ] Others

- [ ] Minimizing False Negative Rate
- [ ] Maximizing Accuracy
- [ ] Maximizing True Positive Rate
- [ ] Minimizing False Positive Rate
- [ ] Minimizing disparity (of FPRs, FNRs, accuracy rates) between demographics

Question 10

1 pts

Please explain your reasoning for the previous question.

HTML Editor

12pt Paragraph

0 words

Quiz saved at 3:15pm Submit Quiz

Quiz (Sept 17th)

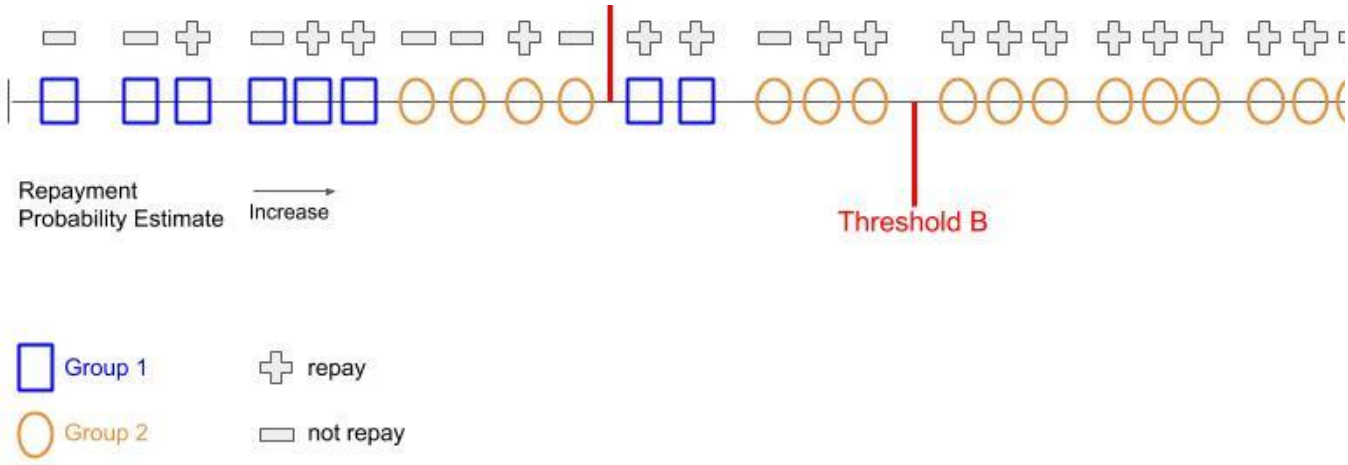
ⓘ This is a preview of the published version of the quiz

Started: Dec 14 at 3:27pm

Quiz Instructions

The first five quiz questions are about designing an ML-based application in a new context: loan application. Some banks use algorithms to make decisions about whether **loan applications** should be granted or denied.

The algorithm learns from historical data to **predict the likelihood that a loan applicant will pay back the loan (or not) and then make a decision based on the likelihood.** The basic logic is that the algorithm considers a loan applicant likely to pay back the loan if his/her profile is similar to profiles of other applicants who have repaid.



Like what we did in the first quiz, let us consider a situation as shown in the figure above. There are two groups (blue square and orange circle). Each individual loan applicant has a label (+ or -) that indicates whether such an individual repays or not. After choosing a threshold, the model will predict “repay” for any individual with a repayment probability estimate higher than the threshold and predict “not repay” for any individual with a repayment probability estimate lower than the threshold.

We will look at the effects on different error rates by changing the threshold.

For sanity check, there are 8 squares and 16 circles in the above picture.

[Concept explanation]
Here we remind you of the definitions of performance metrics.

True positive (TP): people who are being predicted to repay and actually repay
True negative (TN): people who are being predicted to NOT repay and actually do not repay
False positive (FP): people who are being predicted to repay but actually do not repay
False negative (FN): people who are being predicted to not repay but actually repay

True Positive Rate (TPR): The percent of people who are being predicted to repay and actually repay, among all those who actually re-pay. (TP / TP + FN)

True Negative Rate (TNR): The percent of people who are being predicted to NOT repay and actually not repay, among all those who actually do not repay. (TN/ TN + FP)
False Positive Rate (FPR): The percentage of people who are being predicted to re-pay but actually do not repay, among all those who do NOT repay (FP/FP+ TN).
False Negative Rate (FNR): The percentage of people who are being predicted to NOT repay but actually repay, among all those who do repay (FN/FN+TP).

Accuracy: The percentage of loan applicants that are correctly predicted

Disparity: The difference in accuracy/error rates (e.g., false positive rates and false negative rates) between the two groups.

Question 1 — 2 pts

If we pick **threshold A**, calculate the following:

What is the false-positive rate (FPR) for the whole dataset (expressed as a ratio FP/N) : [ ] / [ ]

What is the false-positive rate (FPR) for the Group 1 (blue square) (expressed as a ratio FP_1/N_1): [ ] / [ ]

Question 2 — 1 pts

If we only look at false-positive rates, which of the two groups is benefited?

- ◯ Group 1
- ◯ Both Groups are equally benefited
- ◯ Group 2

Question 3 — 1 pts

Which of the following **reduces false negative rate**?

- ◯ Nothing you can do to minimize the false positive rate.
- ◯ The model chooses a higher threshold.
- ◯ The model chooses a lower threshold.

Question 4

1 pts

Suppose you want to change the threshold to reduce the false-positive rate. How might this affect the false-negative rate?

- ◯ There will be no change in the false negative rate.
- ◯ The false negative rate increases.
- ◯ The false negative rate decrease.

Question 5

1 pts

Suppose you want to change the threshold in the figure to minimize the disparity of false negative rates between two different demographic groups. How might this affect the algorithm’s performance?

- ◯ The accuracy will be lower than the optimal accuracy.
- ◯ The accuracy will still be optimal.
- ◯ The percentage of correct predictions will stay the same.

Question 6

1 pts

Which of the following the current AI/ML technologies are NOT good at? Select all that applies.

- [ ] Understand the world beyond the experience of training
- [ ] Make moral judgement
- [ ] Use common sense
- [ ] Arithmetic calculation

Question 7

1 pts

What are the risks of adding AI/ML predictions into a product or service? Select all that applies.

- [ ] Increase uncertainty
- [ ] Introduce unpredictability
- [ ] Increase speed

No new data to save. Last checked at 3:27pm

Submit Quiz