P6 0:03

Okay, so for those who are watching this recording,

P5 0:07

I was, I was thinking, I was thinking maybe we could first think of like, three questions that we wanted to ask one by one. And then we asked this question and each of us can you explain a little bit of why, like, what are the first metrics that you choose? stuff like that, and then we can explain so that everybody got a chance. What do you think?

P6 0:30

Sure, sure, why don't you start with that? What are what do you choose?

P5 0:34

Okay. Um, I would just repeat what I said. Um, I really look I'm looking for a good accuracy and good false positive rate because I feel um, if you make a mistake on false positive rate is going to have a lot of costs. So that's why choose these two metrics first.

P8 1:01

You guys.

**P6 1:07**

Yeah, I had the same kind of Outlook. So accuracy first, I think that's like a very easy metric to look at. I then looked at the percentage of white to African American to ensure that you know, the, the accuracy was pretty close for both and there wasn't a huge bias or there's always going to be a bias I think that these models but the the lowest bias and then false positive because as you said, That's, I think a more important outcome to have and then false negative is kind of the fourth and then disparity in accuracy. I think I'm still like trying to figure that one out in terms of how it It feeds into the models performance in a way that might sway my

opinion about it.

P7 2:03

I also took the same well and first look at the overview and based on accuracy and false positive rate I ruled out on number two on number eight. And then I like look at this specific for the other ones

P6 2:27

from Europe.

P8 2:34

Okay, so Well, I guess for me, the disparity was the most important thing for some reason, and I'm more interested in low number of false positives. So, you know, this is kind of an unpopular opinion. But I think my first option or Model five, which most of you did not lie, but I chose it because, you know, we knew that the model has false negatives. So possibly you could have another intervention, right? Like to, you know, check the model again to a human in the loop or something like that. And then the next one, I chose model three, like my option because that had high accuracy as well. So those were two of my constraints. That is equality and

no false positives.

P6 3:48

You really isn't. You always have you pronounce?

**P5 3:51**

Yeah, yeah, I do choose five in the beginning. But um, yeah, I think it's the false on negative rate that is making me consider because it hasn't over 90% and is letting a lot of criminal scope. So, um, yeah, in terms of, like the effectiveness of like, machine learning models, you might as well just use human to investigate. So that's why I lower it a little bit. But I do agree with you, you accompany like, I think this is the most ethical model. But yeah, number five, I feel it's the most ethical

model.

**P6 4:40**

Well, the problem with this model is is that you know, you're false. If you have 100, you know, as predictions, right. And you remember in the visualization you saw before that kind of like threshold line if you set the threshold line to only include five people as you're following As your your correct predictions or your your true positives, you're only basing that 1.9% false positive rate on five people, you know, rather than 50%, like if your thresholds right in the middle and you're like, Oh, you know, everyone that I predicted that would not commit a crime didn't except for 1.9%. Like, I feel like that threshold is like super taking into account a super small number of people that have got the answer right for like, you don't know what the whole data set is. Or where that threshold is because it might be super inaccurate, it might say like 95% or 95 people are part of the false negative rate group. Does that make sense? Yeah, you're sampling you might be sampling a small portion and to get that really good false positive rate at 1.9%.

That's what's your model five.

P5 6:00

That makes sense.

P8 6:02

Yeah, that makes sense. Yeah. I mean, I guess my final point was perhaps I mean, you know, you would have another layer, not just a model, right? kind of skeptical about, you know, just the model predicting a criminal. I mean, either case, my head, like, you know, there will be another layer, you know, you have another person checking this process and that person. I mean, basically, the models job is to make that person feel easier. So, if the model was fair, then they know that, okay, fine, it's, it's fair. And in terms of racial groups and other things, all I need to do is check the false negatives piece. But yeah. This is just a subjective opinion. And, yeah, you don't have to really think about these things. focuses on the courts.

P7 7:06

So what are your guys on second option?

P6 7:13

Yeah, I was wondering about these kind of bad more balanced options, or a balanced results of false positive two false negative. Like what that really means like also model four has a very similar FPR to F, nr three.

P8 7:39

Just a high level question. Do you think the color of these images affects the way we pick them? Like, like, if it's green, then your mind is automatically going towards okay. You know, this is good. Ad.

P6 7:57

Yeah, I mean, I picked one in three blue and green. So

P5 8:05

I picked three and seven I think

P7 8:10

I put seven and five.

P5 8:13

I'm gonna list my answers here so that we could see where do we

like agree with each other

P6 9:10

Oh, the disparity in accuracies between white and African. I don't know how I missed that. Sorry guys. I was totally lost on that was

P8 9:27

I think model three. I mean, I thought most number of people

P6 9:34

yeah, model two like it's interesting because you have the same number of are the same accuracy from white to African American like that's that's kind of

P5 9:48

yeah and have really lost and have really low disparity yeah

P8 9:57

yeah.

P7 9:58

But then the pause pause So this is like 45%.

P6 10:04

Well treats everyone fairly, I guess, you know. So I guess we have to put up, put a write up together like which one would be implemented in Allegheny County? I mean, I would say like, three. But then we can also I don't know, do we have kind of a consensus as to which one is or what did we write in the Google Doc? What's the most popular? The

P5 10:31

three is the most popular. Yeah.

P6 10:33

And it looks like I have a second.

P5 10:42

Oh, my God, it's so hard.

P6 10:44

Well, we can we can go i think i think the point of this exercise is to is to realize that each model can serve a different purpose. And there isn't one like overall one model that kind of solves everything. I mean, I think it's important to include Model two, just because it doesn't have any sort of bias. It's it's really like accurate in terms of its accuracy when it relates to race. So maybe we can like in our write up, we can have like two or three kind of responses. And in terms of why we picked those models and what they're good at.

**P5 11:20**

Yeah, I agree. I would also suggest maybe we could write a description for each model. Like for example, we are not saying who's the best but who is like a sorcerer who is like, like a soldier and like, they have different archetypes. And then, while we described them, we could pick the one that we prefer, in some condition, maybe that could show our understanding of all these models.

P8 11:48

Sure.

P5 11:51

You were just talking about model two, which focusing on accuracy

P6 11:58

it would be disparity inaccuracy. So it seems that there wouldn't be

a bias between

you know, accuracy related to race.

But it has bad false positive rates.

P5 12:36

While I'm looking at models one and three, they actually have similar metrics, but we didn't really talk about model one.

Do you think there are there are no significant difference?

P6 13:12

Between which one and which one three and

P5 13:13

one model one and three? Yeah.

P6 13:17

Differences between what? Um,

P5 13:21

yeah, I don't see do do you guys identify any significant difference? Or why you chose moto three? For me? Um, what while looking back I am kind of confused now why I choose three. Maybe it's in the middle.

P6 13:40

Yeah, three and one seems to be very similar. I think that's why I chose both of them as my like one I chose like number three as my one because it's accuracies 1.5% better

show you is that how you pronounce your name? Yeah, got it. Okay, what do you think of model one?

P7 14:11

I was looking at them very closely and I feel like I'm like I'm UAE like UAE said, um, there's really not a huge difference between them The only thing I was like strike me was the disparity in cost native rate. That number one was like 15

and numbers we only not

Unknown Speaker 14:53

yeah

P6 14:56

yeah, I wish I wish we did this part of the exercise before. Did the quiz because now I understand everything way better. Yeah, you're right, though. That's a big difference between that disparity for false negative rate and as a lower false positive rate, though, or disparity.

P5 15:17

Yeah. But that's not the most important metric.

P6 15:21

I know. Cool.

P7 15:26

I feel like no, I'm not sure if I focus too much on false positive rate because I think there's definitely a trade off in like, positive rate and false negative rate. Like, like what we were saying, if you let like, all of the criminals go there, like what's the point? Yeah.

P6 15:46

I think also knowing where that threshold is, like how many results you're basing these on, because it could be like five results that you're basing your false positive rate on because you got one out of five wrong So for one out of 10, so 10%. So it's like also knowing where that threshold is. I don't know if you can tell where the threshold is based on these numbers, but I don't think so.

P5 16:21

So maybe we could talk about In addition, if we will be provided with the information on thresholds, it could better help us make decisions.

Unknown Speaker 16:35

Yeah.

P6 16:47

So I think we should focus on putting that little report together different Are you are you working on that right now?

P8 16:57

You're on mute. Okay. I think you know, But yeah, like you could just synthesize what we have.

If you agree on a model, choose it and buy it. If you are not able to agree, also share your reasons.

Whether you

P6 17:41

think we're saying model three is like the most popular, and then I argue that we should include model two as just a comment on how it would be best for, you know, limiting bias.

But then again like even in model two there's a difference there's like a disparity between the false positive rate and false negative rate. So yeah

P5 19:35

I'm trying to put our descriptions of each model into a table. Oh, do you guys want to talk more about the like model 4567 and maybe why are they not fitting? Yeah

P6 20:10

Four or 545 and six sir Which ones? four, six?

P5 20:16

Yeah, four or 5678 maybe talking about GM

P7 20:25

model I

P8 20:32

mean, do we need to have these in for every model or just have a reason for rejecting all the models?

P5 20:39

Hmm Maybe have a reason for rejecting them?

P6 20:45

Well five is obviously like a huge false negative rate

P7 20:52

90%

P6 20:53

Yeah.

P7 20:55

And then four is a high false positive rate, but then the false positive and false negative rate or more balance.

P5 21:55

Then six would be not doing good in general and all doing bad in like, but like creating equity between

P6 22:05

36 this is model six.

P5 22:07

Yeah, it's has really bad disparity in the arm and fnr Yeah,

P6 22:14

yeah, yeah.

Model six might be the way to go like it's pretty balanced in terms of false positive and false negative. But then

P8 22:35

yeah 20 seconds. Oh,

P6 22:43

okay, let's all present

Transcribed by https://otter.ai