Bias & Fairness

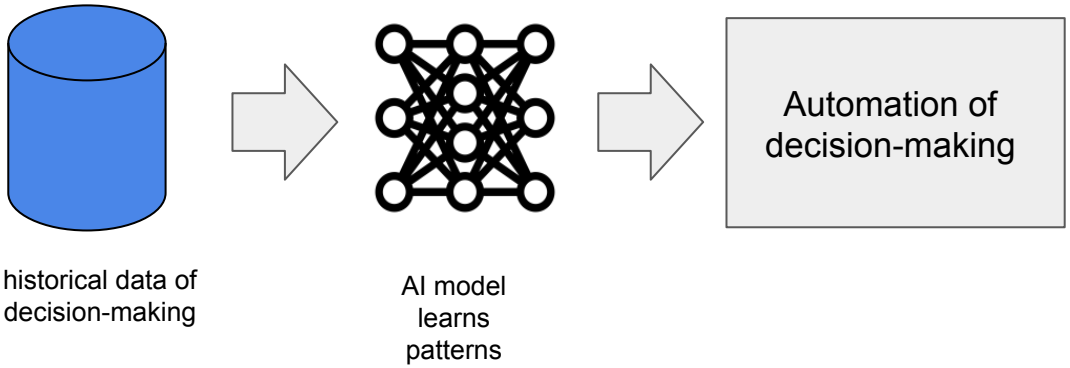
AI in decision Making: Using ML to make yes/no decisions about taking a given action

Examples

- Fraud Prevention
- Credit Lending
- Insurance Policy Pricing
- Preventive Healthcare
- College admission
- Screening Job Candidates

Examples

- Fraud Prevention
- Credit Lending
- Insurance Policy Pricing
- Preventive Healthcare
- College admission
- Screening Job Candidates



Algorithmic decision-making

Can be seen as a statistical risk assessment: we aim to predict the likelihood of an outcome

What is the likelihood of:

- This patients developing disease X?
- This credit card has been stolen?
- This candidate being successful in the job?

We solve this is binary classifiers on historical outcomes

Validation/test sets are used to rank the score (probas, not 0/1) and we pick a threshold for defining 0/1 predictions

Algorithmic decision-making

Can be seen as a statistical risk assessment: we aim to predict the likelihood of an outcome

What is the likelihood of:

- This patients developing disease X?
- This credit card has been stolen?
- This candidate being successful in the job?

We solve this is binary classifiers on historical outcomes

Validation/test sets are used to rank the score (probas, not 0/1) and we pick a threshold for defining 0/1 predictions

Rank	Score	Label
1	0.997	1
2	0.993	1
3	0.986	1
4	0.982	1
5	0.971	0
6	0.965	1
7	0.964	0
8	0.961	0
9	0.953	0
10	0.932	1
11	0.918	0
12	0.873	0
13	0.854	0
14	0.839	0
15	0.777	0
16	0.723	0
17	0.634	0
18	0.512	0
19	0.487	0
20	0.473	0

Total Label Positives: 6
Total Label Negatives: 14
Prevalence: 6/20 = 0.3

Rank	Score	Label
1	0.997	1
2	0.993	1
3	0.986	1
4	0.982	1
5	0.971	0
6	0.965	1
7	0.964	0
8	0.961	0
9	0.953	0
10	0.932	1
11	0.918	0
12	0.873	0
13	0.854	0
14	0.839	0
15	0.777	0
16	0.723	0
17	0.634	0
18	0.512	0
19	0.487	0
20	0.473	0

Predicted Positive: 4

Predicted Negative: 16

Total Label Positives: 6
Total Label Negatives: 14
Prevalence: 6/20 = 0.3

For threshold > 0.980 or top k = 4

True Positives: 4
False Positives : 0
False Negatives : 2
True Negatives : 14

Sources: [1][2][3][4][5][6][7][8][9] view · talk · edit

		Predicted condition			
	Total population = P + N	**Positive (PP)**	**Negative (PN)**	Informedness, bookmaker informedness (BM) = TPR + TNR − 1	Prevalence threshold (PT) $= \frac{\sqrt{TPR \times FPR} - FPR}{TPR - FPR}$
Actual condition	**Positive (P)**	**True positive (TP),** **hit**	**False negative (FN),** **type II error, miss, underestimation**	True positive rate (TPR), recall, sensitivity (SEN), probability of detection, hit rate, power $= \frac{TP}{P} = 1 - FNR$	False negative rate (FNR), miss rate $= \frac{FN}{P} = 1 - TPR$
	Negative (N)	**False positive (FP),** **type I error, false alarm, overestimation**	**True negative (TN),** **correct rejection**	False positive rate (FPR), probability of false alarm, fall-out $= \frac{FP}{N} = 1 - TNR$	True negative rate (TNR), specificity (SPC), selectivity $= \frac{TN}{N} = 1 - FPR$
	Prevalence $= \frac{P}{P+N}$	Positive predictive value (PPV), precision $= \frac{TP}{PP} = 1 - FDR$	False omission rate (FOR) $= \frac{FN}{PN} = 1 - NPV$	Positive likelihood ratio (LR+) $= \frac{TPR}{FPR}$	Negative likelihood ratio (LR−) $= \frac{FNR}{TNR}$
	Accuracy (ACC) $= \frac{TP+TN}{P+N}$	False discovery rate (FDR) $= \frac{FP}{PP} = 1 - PPV$	Negative predictive value (NPV) $= \frac{TN}{PN}$ $= 1 - FOR$	Markedness (MK), deltaP (Δp) = PPV + NPV − 1	Diagnostic odds ratio (DOR) $= \frac{LR+}{LR-}$
	Balanced accuracy (BA) $= \frac{TPR+TNR}{2}$	F_1 score $= \frac{2PPV \times TPR}{PPV + TPR} = \frac{2TP}{2TP + FP + FN}$	Fowlkes–Mallows index (FM) $= \sqrt{PPV \times TPR}$	Matthews correlation coefficient (MCC) $= \sqrt{TPR \times TNR \times PPV \times NPV} - \sqrt{FNR \times FPR \times FOR \times FDR}$	Threat score (TS), critical success index (CSI), Jaccard index $= \frac{TP}{TP + FN + FP}$

source: https://en.wikipedia.org/wiki/Receiver_operating_characteristic

Probabilistic Interpretation

True Positive Rate (Recall) = P(Predicted Positive | Actual Positive)

Positive Predicted Value (PPV) (Precision) = P(Actual Positive | Predicted Positive)

False Positive Rate (FPR) = P(Predicted Positive | Actual Negative)

False Negative Rate (FNR) = P(Predicted Negative | Actual Positive)

False Discovery Rate (FDR) = P(Actual Negative | Predicted Positive) = 1 - PPV

False Omission Rate (FOR) = P(Actual Positive | Predicted Negative)

True Negative Rate (TNR) = P (Predicted Negative | Actual Negative)

Rank	Score	Label
1	0.997	1
2	0.993	1
3	0.986	1
4	0.982	1
5	0.971	0
6	0.965	1
7	0.964	0
8	0.961	0
9	0.953	0
10	0.932	1
11	0.918	0
12	0.873	0
13	0.854	0
14	0.839	0
15	0.777	0
16	0.723	0
17	0.634	0
18	0.512	0
19	0.487	0
20	0.473	0

Predicted Positive: 4

Predicted Negative: 16

Total Label Positives: 6
Total Label Negatives: 14
Prevalence: 6/20 = 0.3

For threshold > 0.980 or top k = 4

True Positives: 4
False Positives : 0
False Negatives : 2
True Negatives : 14

False Positive Rate: 0/14 = 0
Recall: 4/6 = 0.66

False Negative Rate: 2/6 = 0.33
Precision = 4/4 = 1.0

Rank	Score	Label
1	0.997	1
2	0.993	1
3	0.986	1
4	0.982	1
5	0.971	0
6	0.965	1
7	0.964	0
8	0.961	0
9	0.953	0
10	0.932	1
11	0.918	0
12	0.873	0
13	0.854	0
14	0.839	0
15	0.777	0
16	0.723	0
17	0.634	0
18	0.512	0
19	0.487	0
20	0.473	0

Predicted Positive: 10

Predicted Negative: 10

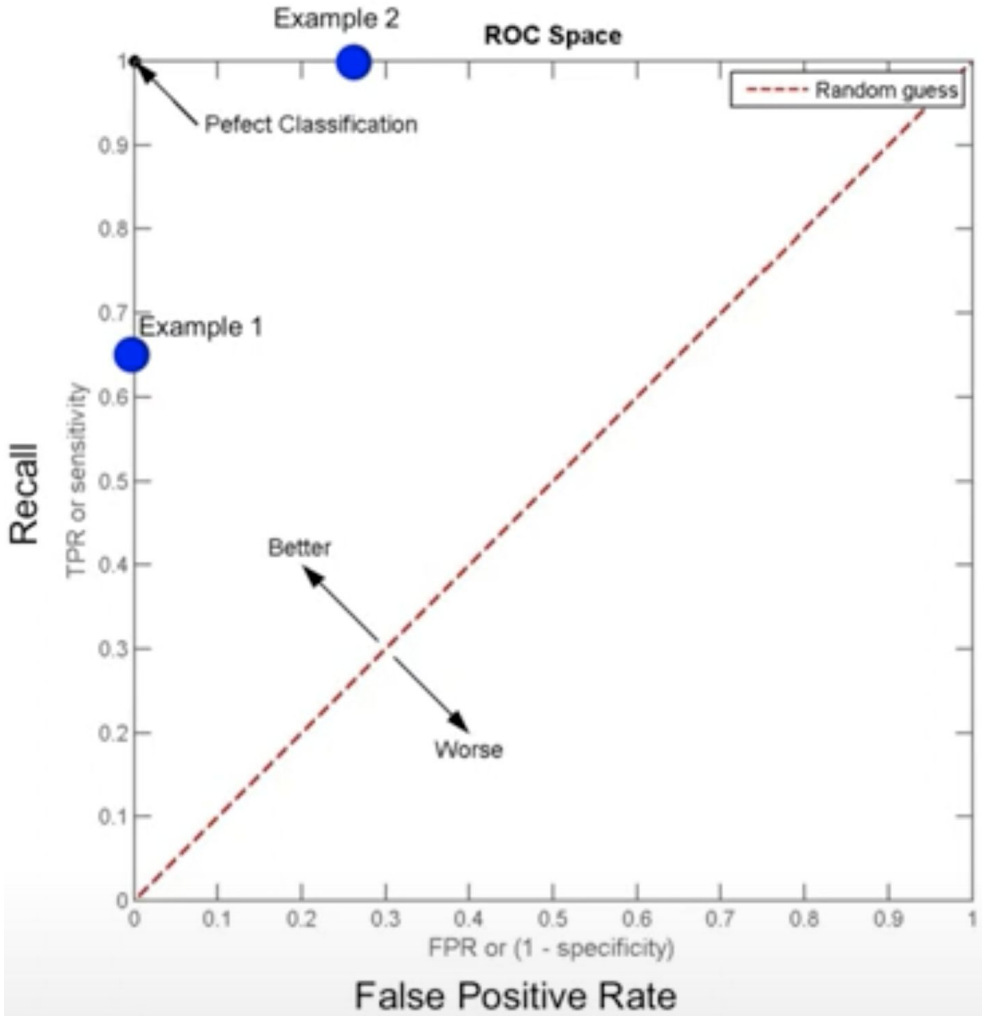
Total Label Positives: 6
Total Label Negatives: 14
Prevalence: 6/20 = 0.3

For threshold > 0.920 or top k = 10

True Positives: 6
False Positives : 4
False Negatives : 0
True Negatives : 10

False Positive Rate: 4/14 = 0.29
Recall: 6/6 = 1.0

False Negative Rate: 0/6 = 0
Precision = 6/10 = 0.6



Recall:
P(y_pred = 1 | label = 1)

FPR:
P(y_pred = 1 | label = 0)

If Recall = FPR, the probability of predicting an instance as positive is independent of the label, therefore it is a random decision

Not all instances are the same…

Rank	Score	Label	Skin Color
1	0.997	1	non-white
2	0.993	1	white
3	0.986	1	white
4	0.982	1	non-white
5	0.971	0	white
6	0.965	1	white
7	0.964	0	white
8	0.961	0	non-white
9	0.953	0	non-white
10	0.932	1	non-white
11	0.918	0	non-white
12	0.873	0	white
13	0.854	0	white
14	0.839	0	white
15	0.777	0	white
16	0.723	0	non-white
17	0.634	0	white
18	0.512	0	non-white
19	0.487	0	white
20	0.473	0	non-white

Overall
Total Label Positives: 6
Total Label Negatives: 14
Prevalence: 6/20 = 0.3

Non-White
Total Label Positives: 3
Total Label Negatives: 6
Group size: 9
Prevalence: 0.33

White
Total Label Positives: 3
Total Label Negatives: 8
Group size: 11
Prevalence: 0.27

Rank	Score	Label	Skin Color
1	0.997	1	non-white
2	0.993	1	white
3	0.986	1	white
4	0.982	1	non-white
5	0.971	0	white
6	0.965	1	white
7	0.964	0	white
8	0.961	0	non-white
9	0.953	0	non-white
10	0.932	1	non-white
11	0.918	0	non-white
12	0.873	0	white
13	0.854	0	white
14	0.839	0	white
15	0.777	0	white
16	0.723	0	non-white
17	0.634	0	white
18	0.512	0	non-white
19	0.487	0	white
20	0.473	0	non-white

Predicted Positive: 10

Predicted Negative: 10

Overall
Total Label Positives: 6
Total Label Negatives: 14
Prevalence: 6/20 = 0.3

Non-White
Total Label Positives: 3
Total Label Negatives: 6
Group size: 9
Prevalence: 0.33

White
Total Label Positives: 3
Total Label Negatives: 8
Group size: 11
Prevalence: 0.27

White

True Positives: 3
False Positives: 2
False Negatives: 0
True Negatives: 6

FPR: 2/8 = **0.25**
Recall: 3/3 = 1.0
Precision: 3/5 = 0.6
FNR = 0

Non-White

True Positives: 3
False Positives: 2
False Negatives: 0
True Negatives: 4

FPR: 2/6 = **0.33**
Recall: 3/3 = 1.0
Precision: 3/5 = 0.6
FNR = 0

Is the previous difference in False Positive Rates between white and non-white a problem?

Is the previous difference in False Positive Rates between white and non-white a problem?

It depends on the action/intervention

Bias in Decision-Making

Decision making is about predictions. There will be errors: false positives and false negatives

Bias is about disparate errors against specific sub-groups.

Decision-Making has been around for thousands of years. Bias as well

Sources of Bias

World

People

Data

- Sample
- Label

Machine Learning Pipeline (Decisions)

Actions

Label Bias

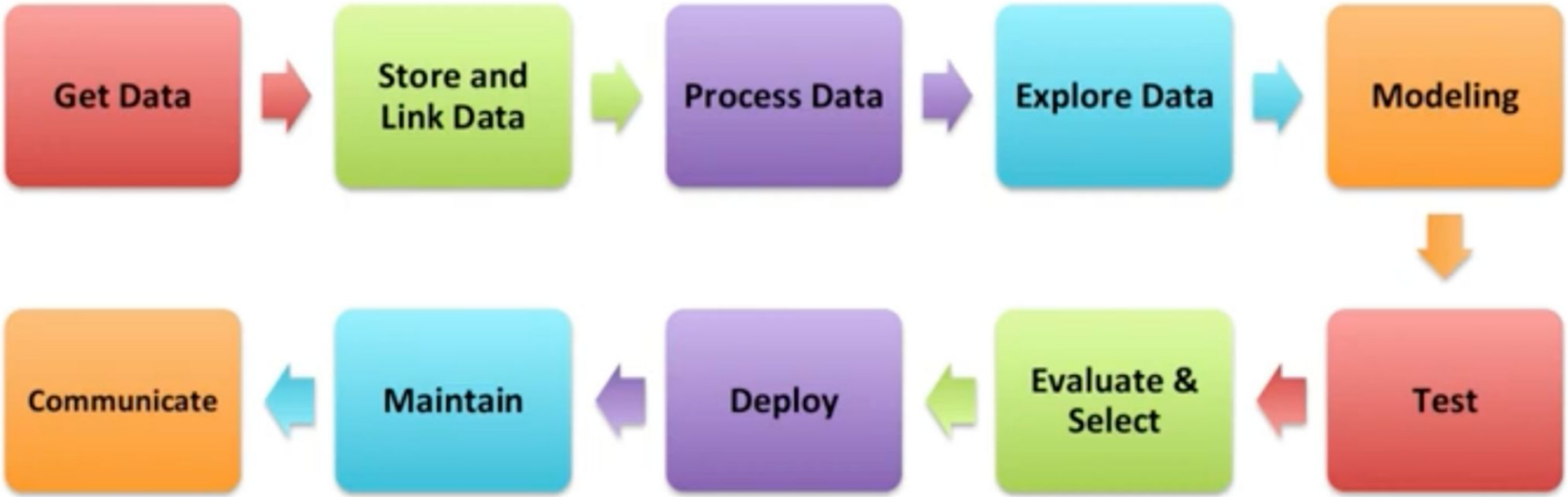
The way the target variable is defined and each data point is labeled might represent disparities between groups.

Differential measurement accuracy across groups (labeling quality).

A variable can be positively correlated with target variable within the majority group but negatively on other groups.

Police Internal Investigations for example.

Bias can be introduced in every step of the ML Pipeline



Action/Intervention Bias

Heterogeneity in the effectiveness of an intervention across groups

Discriminatory ‘overrides’ by the actor conducting an intervention

My fairness definition or yours?

There is no universally-accepted definition of what it means for a decision-making model to be fair.

Punitive Action Example

A model being used to make bail determination (keeping people in jail)

Different people might consider it “fair” if:

It makes mistakes about denying bail to an equal number of white and non-white individuals

Equal count of False Positives

P(wrongly jailed, group i) = C , for all i

Different people might consider it “fair” if:

The chances that a given white or non-white person will be wrongly denied bail is equal, regardless of race

Equal Group Size-Adjusted False Positives

P(wrongly jailed | group i) = C , for all i

Different people might consider it “fair” if:

Among the jailed population, the probability of having been wrongly denied bail is independent of skin color.

Equal False Discovery Rate

P(wrongly jailed | jailed, group i) = C , for all i

Different people might consider it “fair” if:

For people who should be released, the chances that a given white or non-white person will be denied bail is equal

Equal False Positive Rate

P(wrongly jailed | innocent, group i) = C , for all i

Assistive Action Example

A model being used to subsidy diabetes screening and access to preventive care

Different people might consider it “fair” if:

It makes mistakes about denying subsidy to an equal number of women and men.

Equal count of False Negatives

P(missed by benefit, group i) = C , for all i

Different people might consider it “fair” if:

Among the non screened population, the probability of being wrongly denied subsidy is independent of sex

Equal False Omission Rate

P(missed by program | no subsidy, group i) = C , for all i

Different people might consider it “fair” if:

For people who need a social service, the chances that a given woman or man will not get a subsidy is equal.

Equal False Negative Rate

P(missed by subsidy | need assistance, group i) = C , for all i

Parity Measures

Compare a given metric with a reference group

Bias measured as disparity between group metrics

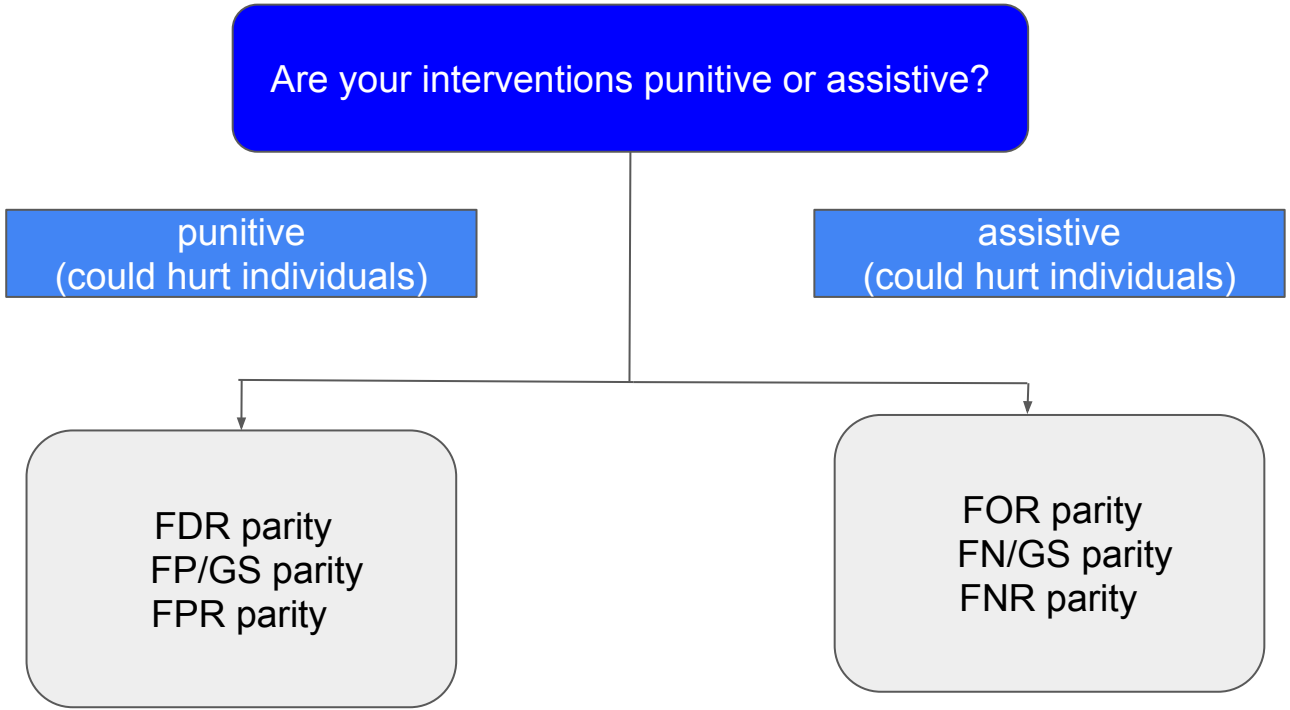
$$FPR_g\ disp = \frac{FPR_{a_i}}{FPR_{a_r}} = \frac{\Pr(\widehat{Y}=1|Y=0,A=a_i)}{\Pr(\widehat{Y}=1|Y=0,A=a_r)}$$

Parity Notion of Fairness

This notion requires that all biases (disparities) be within the range defined by the fairness threshold.

$$\tau \leq DisparityMeasure_{group_i} \leq \frac{1}{\tau}$$

Example: If the fairness threshold is 0.8, the fairness range is between 80% and 125% of the group metric value of the selected reference group."


Are your interventions punitive or assistive?
punitive
(could hurt individuals)
assistive
(could hurt individuals)
FDR parity
FP/GS parity
FPR parity
FOR parity
FN/GS parity
FNR parity

Famous Misconceptions

No “Fairness through Unawareness”

Shall I use race or gender in my models? Remove protected attributes?

Well, other features subsume the protected attributes.

Example: Easy to predict gender based on Facebook likes."

No “Fairness through Demographic Parity”

Example: Accept the same % of women and men for the job as the % of women and men in the population (city? country? candidates pool?)

Decision to be independent from the protected attribute?

Does not ensures “supervised fairness”, as it is possible to have different false positive/negative parities across groups.

Cripples the overall utility metric (e.g. A correlated with Y)

Fairness Tradeoffs

If the base rate (prevalence) is different between groups and the classifier is non-trivial (Recall>0) and imperfect (FPR > 0). Then, either:

- Precision Parity Fails (same as FDR parity)
- FPR and Recall (same as FNR) will be disparate (no equalized odds)

[Kleinberg16, Chouldechova17]