1. Using the Adult dataset from assignment2, we want to judge the fairness of a classifier we have trained on it. In this dataset each instance X is a person, represented through a range of demographic attributes (gender, origin, education, ...). The target variable Y is the income of the person (>50K or <=50K).

Adult dataset (from US in 1994)

- Demographic data
- Income data (target Y)

Dataset:

- 48,000 individuals
- 14 attributes

ID	Feature Name	Feature Type	Feature Values
0	age	continuous	
1	workclass	categorical	Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, ?
2	fnlwgt	continuous	
3	education	categorical	Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool
4	education-num	continuous	
5	marital-status	categorical	Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse
6	occupation	categorical	Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces, ?
7	relationship	categorical	Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried
8	race	categorical	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
9	sex	categorical	Female, Male
10	capital-gain	continuous	
11	capital-loss	continuous	
12	hours-per-week	continuous	
13	native-country	categorical	United-States, Cambodia, England, Puerto-Rico, Canada, Germany, India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, ?

(i) Discuss the following concepts in the context of this data set.

a) Historical Bias

The data set presumably reflects the demographic reality in (parts of) the USA in 1994, however, **society changes!**

In the 1990s presumably fewer women were among the working population; social disparity between different demographic groups was different than today; and education of society in general went up.

=> A classifier trained on the data would **simply reflect the predictive patterns from the past** – a ML model cannot differentiate between “useful knowledge” and “spurious (or unstable) correlations”. It would use all statistical patterns in the data set, and base its predictions on it, and potentially propagate them into the future.

(i) Discuss the following concepts in the context of this data set.

b) Demographic disparity

Machine learning models learn to generalize on the basis of individual observations (labelled training instances). Clearly, the more frequent a model observes a certain pattern (e.g., feature-label combination) the more it will impact the generalisation the model derives.

As such: generalizations are to a large extent **based on the majority classes** in the data set. In the context of the data set above, consider the **"ethnicity (race)" feature**. Presumably, **"white" is the majority class**, and the classifier will make generalizations between education and income largely based on this majority class – however this relation may look very different for different ethnicities.

(i) Discuss the following concepts in the context of this data set.

c) Using the system in the context of (1) a bank which wants to use a model trained on this data for predicting credit ratings; and (2) a government institution in Australia which has access to the features of the Adult for a small population of Australians and wants to predict their income based on it.

Both applications must be viewed with caution, because the function of X (features) → Y (labels) changes between the training scenario and test scenario.

(1) A classifier that's very good at predicting income levels, can probably not be used "off- the-shelf" to predict the credit score of an applicant! Credit worthiness depends on other features as well, and the statistical relation between the given features and the new label may be different. When we deploy a ML algorithm, we must make sure that the task we are solving is the same that the model was trained on.

(2) The **relation** between demographic variables and income are presumably **quite different across countries**, **or continents**! There is little reason to expect that the model trained on the Adult data fares well on a population of Australians. Again: we need to make sure that the data distribution our model observes when it is deployed is (as) similar (as possible) to the data distribution it was trained on.

(ii) You are asked to develop an income classifier that is fair with respect to the protected attribute *gender*. Your boss is a big believer in logistic regression classifiers, and asks you to apply this particular classifier architecture with no modification. What approach(es) could you take to still test/improve the performance of your classifier?

The lectures covered three approaches to improving fairness of classifiers: **(1) changing the data; (2) changing the ML model/loss; (3) post-processing of the model predictions.** Our boss prohibited option (2), so options (1) and (3) are left.

For option (1), **pre-processing the data**, we could re-sample the data set such that both groups are represented similarly in the data set, for example by down-sampling instances of the majority class. In this way, our classifier would base its generalisations on both groups, rather than focussing on the majority group disproportionately.

Similarly, we could assign each instance a weight and penalize model errors for instances with higher weight more than instances with lower weight. We would compare the **true probability** of observing a label with a protected group against the ***expected*** **probability** if the two were independent.

(ii) You are asked to develop an income classifier that is fair with respect to the protected attribute *gender*. Your boss is a big believer in logistic regression classifiers, and asks you to apply this particular classifier architecture with no modification. What approach(es) could you take to still test/improve the performance of your classifier?

Pre-processing

Expected distribution (if $A \perp Y$)

$$P_{exp}(A{=}a, Y{=}1) = P(A{=}a) \times P(Y{=}1) = \frac{\#(A{=}a)}{|D|} \times \frac{\#(Y{=}1)}{|D|}$$

Observed distribution

$$P_{obs}(A{=}a, Y{=}1) = \frac{\#(Y{=}1, A{=}a)}{|D|}$$

Weigh each instance by

$$W(X_i{=}\{x_i, a_i, y_i\}) = \frac{P_{exp}(A{=}a_i, Y{=}y_i)}{P_{obs}(A{=}a_i, Y{=}y + i)}$$

(ii) You are asked to develop an income classifier that is fair with respect to the protected attribute *gender*. Your boss is a big believer in logistic regression classifiers, and asks you to apply this particular classifier architecture with no modification. What approach(es) could you take to still test/improve the performance of your classifier?

(2) we could leave the data untouched, and instead **post-process** the classifier output.

Rather than applying a decision threshold of 0.5 ($\hat{y}$=1 if score >0.5 and $\hat{y}$=0 if score <=0.5) we could devise a group-specific threshold for female and male applicants. This threshold could for example be set such that for both genders we expect a comparable number of false negative predictions under our already trained classifier (which corresponds to ***equal opportunity***).

(ii) You are asked to develop an income classifier that is fair with respect to the protected attribute *gender*. Your boss is a big believer in logistic regression classifiers, and asks you to apply this particular classifier architecture with no modification. What approach(es) could you take to still test/improve the performance of your classifier?

Post-processing

Modify the classifier predictions (scores s or labels $\hat{y}$

- E.g., decide on individual thresholds per group, such that:

$$\hat{y}_i = 1 \text{ if } s_i > \theta_i$$

- Come up with a special strategy for "difficult" instances, i.e., instances where $P(\hat{y}_i) \approx 0.5$

2. Using the dataset in question 1, assume we selected *gender* as our protected attribute. We trained our classifier and observed the following outcomes.

$P(\hat{y}=1\|A=f)$	$P(\hat{y}=1\|A=m)$	$P(\hat{y}=1\|Y=1, A=f)$	$P(\hat{y}=1\|Y=1, A=m)$	$P(Y=1\|\hat{y}=1, A=f)$	$P(Y=1\|\hat{y}=1, A=m)$	$P(Y=1\|\hat{y}=1)$	$P(\hat{y}=1\| Y=1)$
0.81	0.75	0.80	0.86	0.73	0.74	0.74	0.85

(i). Name each of the statistics and provide a formula for its measurement. Be sure you understand the intuition / connection behind the statistical notion and its metric.

- **Positive rate:** fraction of (true or false) positives predicted for each group
- **True Positive rate (TPR):** fraction of positives among all positives in the data (**recall**)
- **Positive predictive value (PPV):** fraction of true positives among all positive predictions (**precision**)

All of these can be computed for individual groups (hence the conditioning) or for the overall population (last two columns)

2. Using the dataset in question 1, assume we selected *gender* as our protected attribute. We trained our classifier and observed the following outcomes.

P(ŷ=1|A=f)	P(ŷ=1|A=m)	P(ŷ=1|Y=1, A=f)	P(ŷ=1|Y=1, A=m)	P(Y=1|ŷ=1, A=f)	P(Y=1|ŷ=1, A=m)	P(Y=1|ŷ=1)	P(ŷ=1| Y=1)
0.81	0.75	0.80	0.86	0.73	0.74	0.74	0.85

(i). Name each of the statistics and provide a formula for its measurement. Be sure you understand the intuition / connection behind the statistical notion and its metric.

P(ŷ=1|A=f)	P(ŷ=1|A=m)
0.81	0.75
Female Positive Rate	Male Positive Rate
$PR_f = \frac{P}{N_f}$	$PR_m = \frac{P}{N_m}$

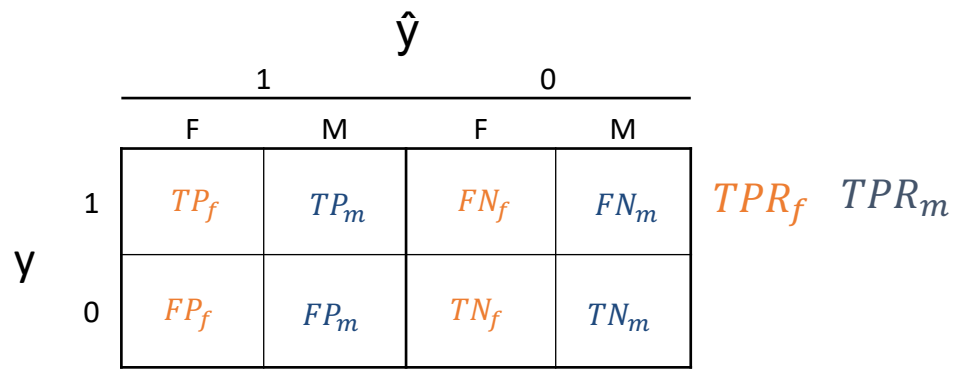
		ŷ			
		1		0	
		F	M	F	M
y	1	TP_f	TP_m	FN_f	FN_m
	0	FP_f	FP_m	TN_f	TN_m

2. Using the dataset in question 1, assume we selected *gender* as our protected attribute. We trained our classifier and observed the following outcomes.

P(ŷ=1|A=f)	P(ŷ=1|A=m)	P(ŷ=1|Y=1, A=f)	P(ŷ=1|Y=1, A=m)	P(Y=1|ŷ=1, A=f)	P(Y=1|ŷ=1, A=m)	P(Y=1|ŷ=1)	P(ŷ=1| Y=1)
0.81	0.75	0.80	0.86	0.73	0.74	0.74	0.85

(i). Name each of the statistics and provide a formula for its measurement. Be sure you understand the intuition / connection behind the statistical notion and its metric.

P(ŷ=1|Y=1, A=f)	P(ŷ=1|Y=1, A=m)
0.80	0.86
Female True Positive Rate (Recall)	Male True Positive Rate (Recall)
$TPR_f = \frac{TP_f}{TP_f + FN_f}$	$TPR_m = \frac{TP_m}{TP_m + FN_m}$

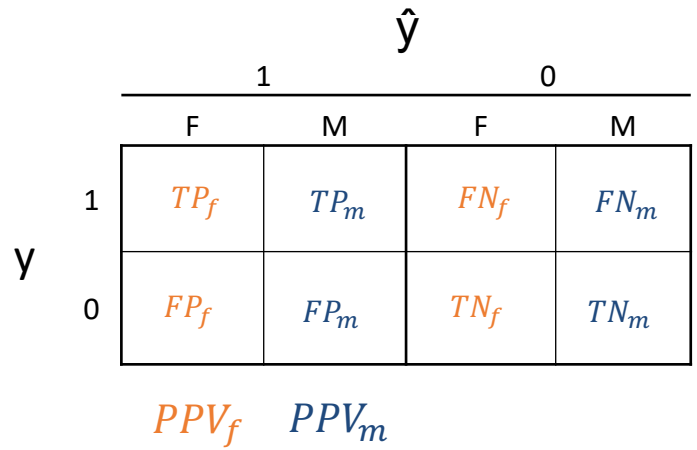


2. Using the dataset in question 1, assume we selected *gender* as our protected attribute. We trained our classifier and observed the following outcomes.

P(ŷ=1|A=f)	P(ŷ=1|A=m)	P(ŷ=1|Y=1, A=f)	P(ŷ=1|Y=1, A=m)	P(Y=1|ŷ=1, A=f)	P(Y=1|ŷ=1, A=m)	P(Y=1|ŷ=1)	P(ŷ=1| Y=1)
0.81	0.75	0.80	0.86	0.73	0.74	0.74	0.85

(i). Name each of the statistics and provide a formula for its measurement. Be sure you understand the intuition / connection behind the statistical notion and its metric.

P(Y=1|ŷ=1, A=f)	P(Y=1|ŷ=1, A=m)
0.73	0.74
Female True Positive Predictive Value (**precision**)	Male True Positive Predictive Value (**precision**)
$PPV_f = \frac{TP_f}{TP_f + FP_f}$	$PPV_m = \frac{TP_m}{TP_m + FP_m}$

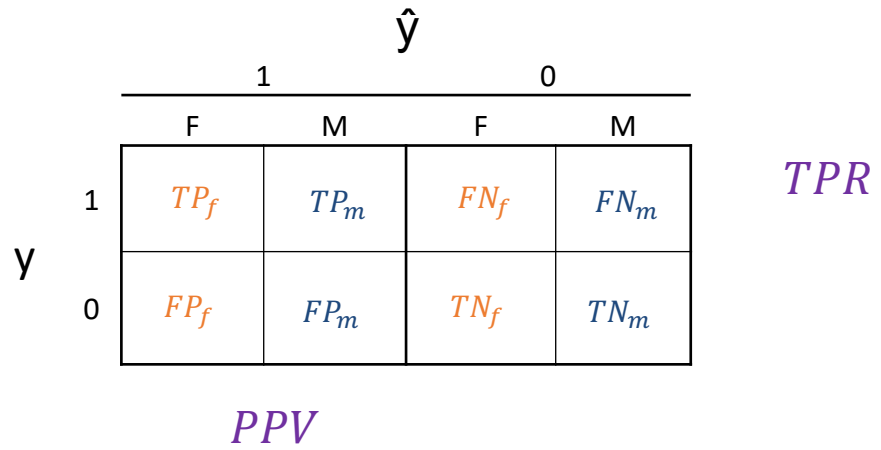


2. Using the dataset in question 1, assume we selected *gender* as our protected attribute. We trained our classifier and observed the following outcomes.

P(ŷ=1|A=f)	P(ŷ=1|A=m)	P(ŷ=1|Y=1, A=f)	P(ŷ=1|Y=1, A=m)	P(Y=1|ŷ=1, A=f)	P(Y=1|ŷ=1, A=m)	P(Y=1|ŷ=1)	P(ŷ=1| Y=1)
0.81	0.75	0.80	0.86	0.73	0.74	0.74	0.85

(i). Name each of the statistics and provide a formula for its measurement. Be sure you understand the intuition / connection behind the statistical notion and its metric.

P(ŷ=1| Y=1)	P(Y=1|ŷ=1)
0.85	0.74
Recall: True Positive Rate (overall)	**Precision:** Predictive Parity Value (overall)
$TPR = \frac{TP}{TP + FN}$	$PPV = \frac{TP}{TP + FP}$



(ii). For each of the following criteria, decide whether the classifier meets this criterion.

a) Group Fairness (Demographic parity)

$P(\hat{y}=1\|A=f)$	$P(\hat{y}=1\|A=m)$
0.81	0.75
Female Positive Rate	Male Positive Rate
$PR_f = \frac{P}{N_f}$	$PR_m = \frac{P}{N_m}$

Demographic parity requires that the **positive rate of all groups is identical** (i.e., the first two columns in our case).

There is a substantial gap: the fraction of positive predictions for females (0.81) is **larger** than the fraction of positive predictions for males (0.75). Our classifier does not achieve group fairness.

(ii). For each of the following criteria, decide whether the classifier meets this criterion.

b) Equal opportunity

$P(\hat{y}=1 \mid Y=1, A=f)$	$P(\hat{y}=1 \mid Y=1, A=m)$
0.80	0.86
Female True positive rate (**Recall**)	Male True positive rate (**Recall**)
$TPR_f = \frac{TP_f}{TP_f + FN_f}$	$TPR_m = \frac{TP_m}{TP_m + FN_m}$

Equal opportunity requires that the **true positive rates are identical across groups**:

Intuitively, if we know that an individual is credit worthy, our classifier should predict the individual as credit worthy with the same probability – irrespective of the gender of the applicant.

We compare columns 3 and 4 in the table above and find that our classifier does not achieve equal opportunity: the TPR is higher for males (0.86) than for females (0.80).

This means that our classifier is more likely to correctly grant credit to a man (who indeed is credit worthy) than to a woman (who indeed is credit worthy).

Or, equivalently: our classifier is more likely to falsely deny a credit to a woman (who indeed is credit worthy) than to a man (who indeed is credit worthy). This is because False Negative Rate is also different across the group. We calculate FNR as

FNR = 1-TPR

(ii). For each of the following criteria, decide whether the classifier meets this criterion.

c) Predictive Parity

P(Y=1|ŷ=1, A=f)	P(Y=1|ŷ=1, A=m)
0.73	0.74
Female True Positive Predictive Value (**precision**)	Male True Positive Predictive Value (**precision**)
$PPV_f = \frac{TP_f}{TP_f + FP_f}$	$PPV_m = \frac{TP_m}{TP_m + FP_m}$

Predictive parity requires that **positive predictive values are identical across groups**.

Intuitively, if we have predicted a positive rating for an applicant, it should be equally likely that the applicant is indeed credit worthy – irrespective of the gender.

We compare columns 5 and 6 in the table above and find that both values are similar (if not identical). Although strictly speaking the classifier doesn't exhibit perfect predictive parity, for any realistic application reducing differences to below a certain threshold is typically sufficient.

For almost all scenarios, PPV values of 0.73 and 0.74 for male and female would be considered fair. Our classifier does achieve predictive parity.☺️

3. A common metric for assessing classifier fairness is the **GAP** in scores achieved across groups. If we choose true positive rate (TPR) as our score of interest, we will check the classifier for "equal opportunity". If we choose positive predictive value as score of interest, we test our classifier for "predictive parity". Verify your observations in question 2 using (a) max-GAP and (b) avg-GAP. When would avg-GAP be preferred, and when max-GAP?

We compute the average across all groups g, as following where ϕ_g denotes a group-specific score (TPR or PPV) and ϕ denotes the overall value.

$$GAP_{avg} = \frac{1}{G}\sum_{g=1}^{G}|\phi_g - \phi|$$

$$GAP_{max} = max_{g \in G}|\phi_g - \phi|$$

If we choose true positive rate (TPR) as our score of interest, we will check the classifier for “equal opportunity”.

TPR:

$$GAP_{avg} = \frac{1}{2}(|0.8 - 0.85| + |0.86 - 0.85|) = \frac{1}{2}(0.5 + 0.01) = 0.03$$
$$GAP_{max} = max(|0.8 - 0.85|, |0.86 - 0.85|) = max(0.5, 0.01) = 0.5$$

$P(\hat{y}=1\|Y=1, A=f)$	$P(\hat{y}=1\|Y=1, A=m)$	$P(\hat{y}=1\|Y=1)$
0.80	0.86	0.85
Female True Positive Rate (Recall)	Male True positive rate (Recall)	**Recall**: True Positive Rate (overall)
$TPR_f = \frac{TP_f}{TP_f + FN_f}$	$TPR_m = \frac{TP_m}{TP_m + FN_m}$	$TPR = \frac{TP}{TP + FN}$

If we choose positive predictive value as score of interest, we test our classifier for "predictive parity".

PPV:

$$GAP_{avg} = \frac{1}{2}(|0.73 - 0.74| + |0.74 - 0.74|) = \frac{1}{2}(0.01 + 0.00) = 0.005$$
$$GAP_{max} = max(|0.73 - 0.74|, |0.74 - 0.74|) = max(0.01,0.00) = 0.01$$

P(Y=1|ŷ=1, A=f)	P(Y=1|ŷ=1, A=m)	P(Y=1|ŷ=1)
0.73	0.74	0.74
Female True Positive Predictive Value (**precision**)	Male True Positive Predictive Value (**precision**)	**Precision**: Predictive Parity Value (overall)
$PPV_f = \frac{TP_f}{TP_f + FP_f}$	$PPV_m = \frac{TP_m}{TP_m + FP_m}$	$PPV = \frac{TP}{TP + FP}$

When would avg-GAP be preferred, and when max-GAP?

Think about a scenario where we have 5 different sensitive groups (e.g. different ethnicities), and it is very important that no single ethnicity is treated unfairly.

The **avg-GAP** could look reasonable if some ethnicities achieve far larger TPR than the overall measure and others achieve much lower TPR.

The **max-GAP** on the other hand would capture every individual outlier.

So: in situations where **outliers are really unacceptable**, **max-GAP** should be used.

4. For our classifier above, we reported that TPR_f=0.8, TPR_m=0.86 and TPR=0.85. How do you think TPR was computed, and what does it tell us about the data?

P(ŷ=1|Y=1, A=f)	P(ŷ=1|Y=1, A=m)	P(ŷ=1| Y=1)
0.80	0.86	0.85
Female True Positive Rate (Recall)	Male True Positive Rate (Recall)	**Recall**: True Positive Rate (overall)
$TPR_f = \frac{TP_f}{TP_f + FN_f}$	$TPR_m = \frac{TP_m}{TP_m + FN_m}$	$TPR = \frac{TP}{TP + FN}$

4. For our classifier above, we reported that TPR_f=0.8, TPR_m=0.86 and TPR=0.85. How do you think TPR was computed, and what does it tell us about the data?

macro-averaging: calculate P, R per class and then average

$$\text{Precision}_M = \frac{\sum_{i=1}^{c} \text{Precision}_i}{c}$$

$$\text{Recall}_M = \frac{\sum_{i=1}^{c} \text{Recall}_i}{c}$$

micro-averaging: combine all test instances into a single pool

$$\text{Precision}_\mu = \frac{\sum_{i=1}^{c} TP_i}{\sum_{i=1}^{c} TP_i + FP_i}$$

$$\text{Recall}_\mu = \frac{\sum_{i=1}^{c} TP_i}{\sum_{i=1}^{c} TP_i + FN_i}$$

P(ŷ=1|Y=1, A=f)	P(ŷ=1|Y=1, A=m)	P(ŷ=1| Y=1)
0.80	0.86	0.85
Female True Positive Rate (Recall)	Male True Positive Rate (Recall)	**Recall**: True Positive Rate (overall)
$TPR_f = \frac{TP_f}{TP_f + FN_f}$	$TPR_m = \frac{TP_m}{TP_m + FN_m}$	$TPR = \frac{TP}{TP + FN}$