Artificial Intelligence Ethics: 4th lesson – Artificial Intelligence Fairness

There are many different ways of defining what we might look for in a fair machine learning (ML) model. For instance, say we're working with a model that approves (or denies) credit card applications. Is it fair if the approval rate is equal across genders, or better if gender classification is removed from the dataset and hidden from the model?

In this tutorial, we'll cover four criteria that we can use to decide if a model is fair. Then, you'll apply what you've learned in a hands-on exercise where you'll run code to train several models and analyze fairness of each.

Criteria of AI fairness:

These four fairness criteria are a useful starting point, but it's important to note that there are more ways of formalizing fairness, which you are encouraged to explore. Assume we're working with a model that selects individuals to receive some outcome. For instance, the model could select people who should be approved for a loan, accepted to a university, or offered a job opportunity (so we don't consider models that perform tasks like facial recognition or text generation, among other things).

* ***Demographic parity***

Demographic parity says the model is fair if the composition of people who are selected by the model matches the group membership percentages of the applicants.

* ***Equal opportunity***

Equal opportunity fairness ensures that the proportion of people who should be selected by the model ("positives") that are correctly selected by the model is the same for each group. We refer to this proportion as the true positive rate (TPR) or sensitivity of the model.

* ***Equal accuracy***

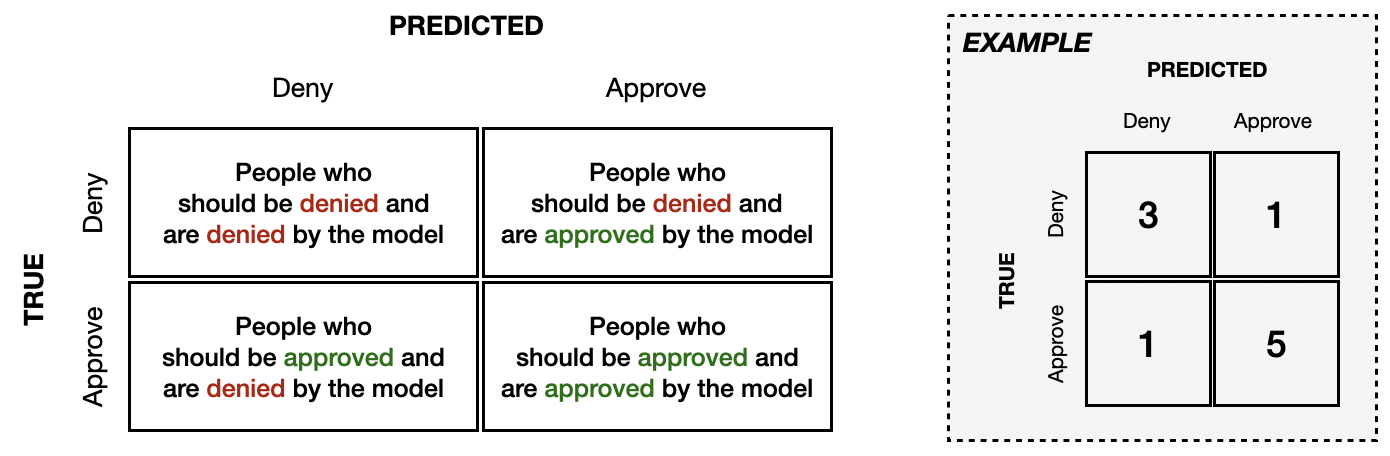
Alternatively, we could check that the model has equal accuracy for each group. That is, the percentage of correct classifications (people who should be denied and are denied, and people who should be approved who are approved) should be the same for each group. If the model is 98% accurate for individuals in one group, it should be 98% accurate for other groups.

* ***Group unawareness***

Group unaware fairness removes all group membership information from the dataset. For instance, we can remove gender data to try to make the model fair to different gender groups. Similarly, we can remove information about race or age.

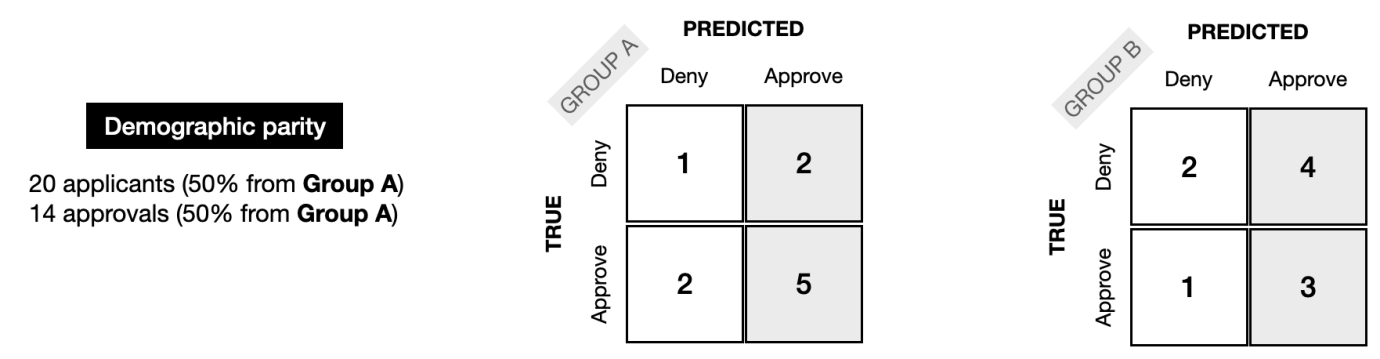
Case study example:

We'll work with a small example to illustrate the differences between the four different types of fairness. We'll use a confusion matrix, which is a common tool used to understand the performance of a ML model. This tool is depicted in the example below, which depicts a model with 80% accuracy (since 8/10 people were correctly classified) and has an 83% true positive rate (since 5/6 "positives" were correctly classified).

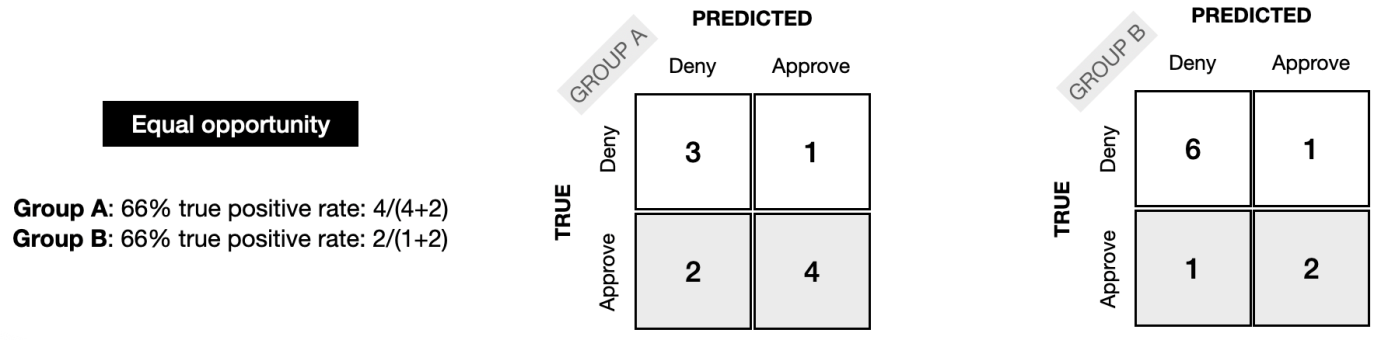


To understand how a model's performance varies across groups, we can construct a different confusion matrix for each group. In this small example, we'll assume that we have data from only 20 people, equally split between two groups (10 from Group A, and 10 from Group B).

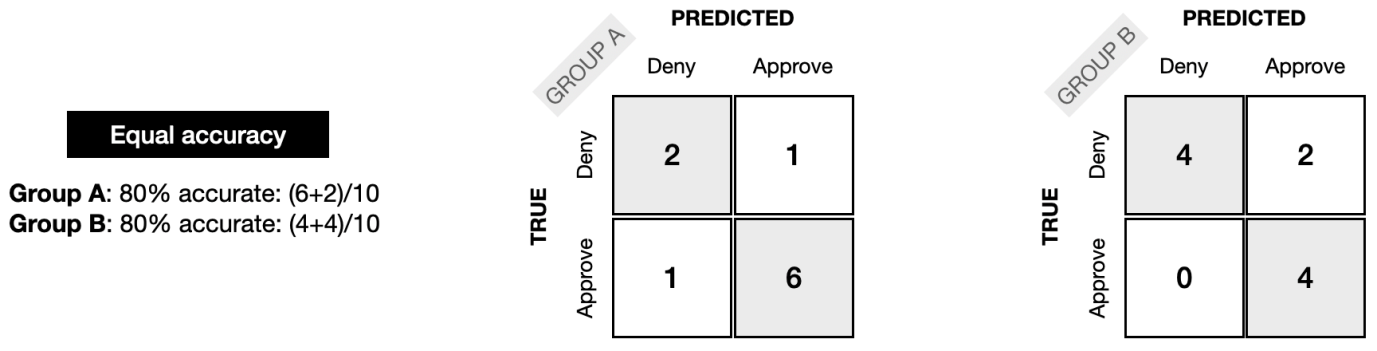
The next image shows what the confusion matrices could look like, if the model satisfies demographic parity fairness. 10 people from each group (50% from Group A, and 50% from Group B) were considered by the model. 14 people, also equally split across groups (50% from Group A, and 50% from Group B) were approved by the model.



For equal opportunity fairness, the TPR for each group should be the same; in the example below, it is 66% in each case.



Next, we can see how the confusion matrices might look for equal accuracy fairness. For each group, the model was 80% accurate.



Note that group unaware fairness cannot be detected from the confusion matrix, and is more concerned with removing group membership information from the dataset.

Take the time now to study these toy examples, and use it to build your intuition for the differences between the different types of fairness. How does the example change if Group A has double the number of applicants of Group B?

Also note that none of the examples satisfy more than one type of fairness. For instance, the demographic parity example does not satisfy equal accuracy or equal opportunity. Take the time to verify this now. In practice, it is not possible to optimize a model for more than one type of fairness: to read more about this, explore the *Impossibility Theorem of Machine Fairness*. So which fairness criterion should you select, if you can only satisfy one? As with most ethical questions, the correct answer is usually not straightforward, and picking a criterion should be a long conversation involving everyone on your team.

When working with a real project, the data will be much, much larger. In this case, confusion matrices are still a useful tool for analyzing model performance. One important thing to note, however, is that real-world models typically cannot be expected to satisfy any fairness definition perfectly. For instance, if "demographic parity" is chosen as the fairness metric, where the goal is for a model to select 50% men, it may be the case that the final model ultimately selects some percentage close to, but not exactly 50% (like 48% or 53%).