An Introduction to Fairness in Machine Learning

Rahul Shekhar · Follow
Published in Analytics Vidhya
9 min read · Aug 19, 2020

Listen Share More

The Goal

attempts to uncover patterns in this data to master the task at hand. However, no EXPLICIT instructions are given to model as to what patterns it should or should not look for. As a direct consequence, many a time, the model ends up learning unwanted hypotheses. Concretely, data-driven decision making is only as reliable as the data on which it is based.

Machine Learning models are heavily ingrained in sensitive decision making processes such as hiring decisions, loan approvals, selecting prospective clients etc. It is of utmost importance that these models are not biased and treat everyone fairly. For instance, data from the National Center for Women & Information Technology (NCWIT) shows that only 26% of the computing occupations are held by women. Furthermore, only African American women only represent 3% of that number. Clearly, models based on this data will learn these biases and penalize minorities unnecessarily, despite being qualified for the role. Thus, the significance of embracing techniques to measure and remove biases from machine learning models cannot be overstated.

I've read a decent number of articles on methods to measure bias present in machine learning models. However, none of them were able to explain the concepts in an intuitive manner that was easy to grasp. This motivated me to write this article, in the hope of breaking down some of these concepts and encouraging more interest in this area, as it is going to be very prevalent in the near future. We really don't want machine learning to perpetuate our social prejudices — something good turning evil!

Protected/Sensitive Attributes

Following are some of the most common attributes to evaluate bias on:

- Gender
- Ethnicity
- Religion
- Age
- Marital Status

- Citizenship

These are known as sensitive or protected attributes because one's machine learning model should not be biased on them. Depending upon the features you chose to include in your model, there could be many more. This directly takes us into the next important concept of PROXY attributes. Does simply removing these sensitive attributes from your data before training make the model unbiased? Sadly, the answer is NO. This is because certain attributes implicitly contain the biases present in these sensitive attributes. For instance, it has been shown that neighborhood and zip code carry the same biases as in that of ethnicity, making them proxy attributes of ethnicity. Furthermore, depending on the problem at hand, these attributes cannot always be removed. Hence, one always has to conduct fairness tests in order to verify that their model is unbiased.

Measuring Bias

Okay, now for the intriguing stuff! How do we actually investigate whether a model is biased or not? There are three criteria to measure this — Independence, Separation and Sufficiency. According to me, the best way to understand these criteria is by using an intuitive example rather than trying to wrap one's head around a general formula, especially one with conditional and independence symbols (yeesh!). Given that I am a recent graduate, I have decided to use college admission decisions as the basis of my example. Below is a table I generated in order to showcase some quick calculations.

S. No	Gender	True Label	Predicted Label
1	M	1	1
2	M	1	1
3	F	0	0
4	M	0	1
5	F	1	0
6	M	1	0
7	M	1	1
8	F	1	1
9	M	0	0
10	F	0	0

The sensitive attribute of choice is gender and can take on the value male and female. The label 1 refers to “You were good enough! “ and 0 refers to “You were good enough, however .. ”. This model is trying to replicate the applicant screening process and thus, the prediction labels depict who would have been accepted or rejected if we were to replace the current process with this model.

I will also state the equations to satisfy each of these criteria for a binary sensitive attribute which can easily be extended. The following notation will be used:

- **A** sensitive/protected attribute
- **C**=C(X, A) is the learned classifier
- **Y** target variable

Independence - Equality of outcomes/selection

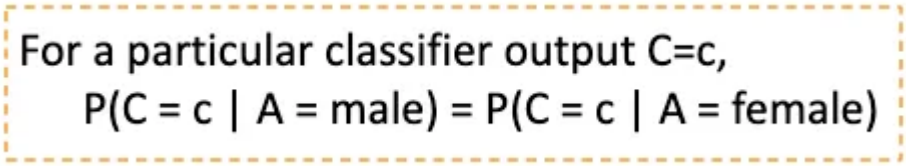
Example: The acceptance rate of males and females should be the same!

Looking at the above example, we can calculate that the acceptance rate of males is 4/6 = 66.67%, while that of females is 1/4 = 25% (calculated using the predicted label). However, in order to satisfy the independence criteria the selection rate of both males and females should be the same. Furthermore, if we calculate the acceptance rate of the data used for training the model, the acceptance rate of males is 4/6 = 66.67% and that of females is 2/4 = 50%. In other words, to satisfy the

independence criteria, most times we can't achieve a perfect model (a model that has 100% accuracy) because the training data doesn't meet the criteria.

Statistically speaking, the classification scores should be independent of the sensitive attribute. Another way to interpret this is that there should be 0 mutual information between the classifier scores and the sensitive attribute.

General Formula,



Separation - Equality of Errors (Equality of outcomes given a threshold)

Example: The rejection rate of males and females DESPITE being qualified enough for admission (false negative) should be the same!

This is a slight relaxation of the separation rule because it doesn't mention that the false positive rate should be the same as well, however it gets the crux of the concept across. This criteria promotes the concept that "similar people should be treated similarly" i.e once we have a set of candidates that we know are qualified enough for admission there should not be any bias when selecting a subset of these candidates for admission.

It is useful to think of separation as a measure of equality of errors. In other words, the chances of predicting a false positive and a false negative of each group should be the same. In the above data, the false negative rate for males = 1/4 = 25% and the false negative rate for females is 1/2 = 50%. Thus, this model does not meet the separation criteria.

One advantage of this criterion is that it is still possible to learn a perfect model. If a model has 100% accuracy, it means that the model satisfies the separation criterion.

General Formula,

For a particular classifier output C=c,

$$P(C = c \mid A = male, Y = True) = P(C = c \mid A = female, Y = True)$$
$$P(C = c \mid A = male, Y = False) = P(C = c \mid A = female, Y = False)$$

One of the reasons that Separation might be more desirable than Independence is because there might be some correlation between the sensitive characteristic and target variable. For example, a certain company might state that based on certain measures — CLV (Customer Lifetime Value), Default Rate etc — that having a different rate of choosing clients from different groups is a business necessity (cannot meet the independence criteria). However, separation only allows correlation to the extent of the target variable i.e there should be no biases within the group of people that should be chosen as clients.

Sufficiency - Choices reflect same accuracy per group (calibration)

Example: The chances of males and females being qualified enough given the admission decision (predicted variable) should be the same!

This criteria ensures that the sensitive attribute and the target variable Y are clear from context. In other words, given a score the probability of the true variable being 1 should be the same for each group. In our example, this means the prediction returned for whether a candidate should be accepted or not reflects the candidate's qualifications accurately. If it is the same for male and females then we can safely say that our model is able to correctly learn what comprises a good candidate without biasing on gender.

This criterion also has the possibility of a perfect model being learned. We will introduce the sufficiency formula before doing any calculations because sufficiency is less intuitive than the other two conditions.

For a particular classifier output C=c,

$$P(Y = True \mid A = male, C = c) = P(Y = True \mid A = female, C = c)$$

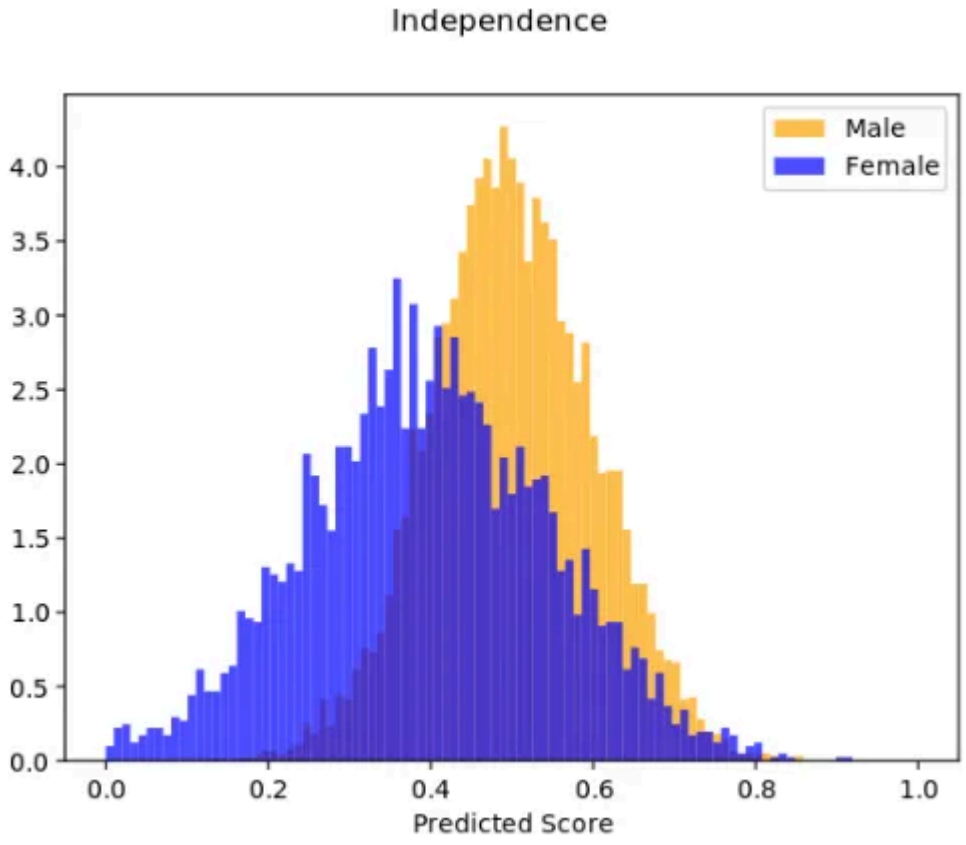
In the example data,

P(Y = 1 | c=1, A = Male) = ¾ = 75%; P(Y=1 | c=1, A = Female) = 1/1 = 100%

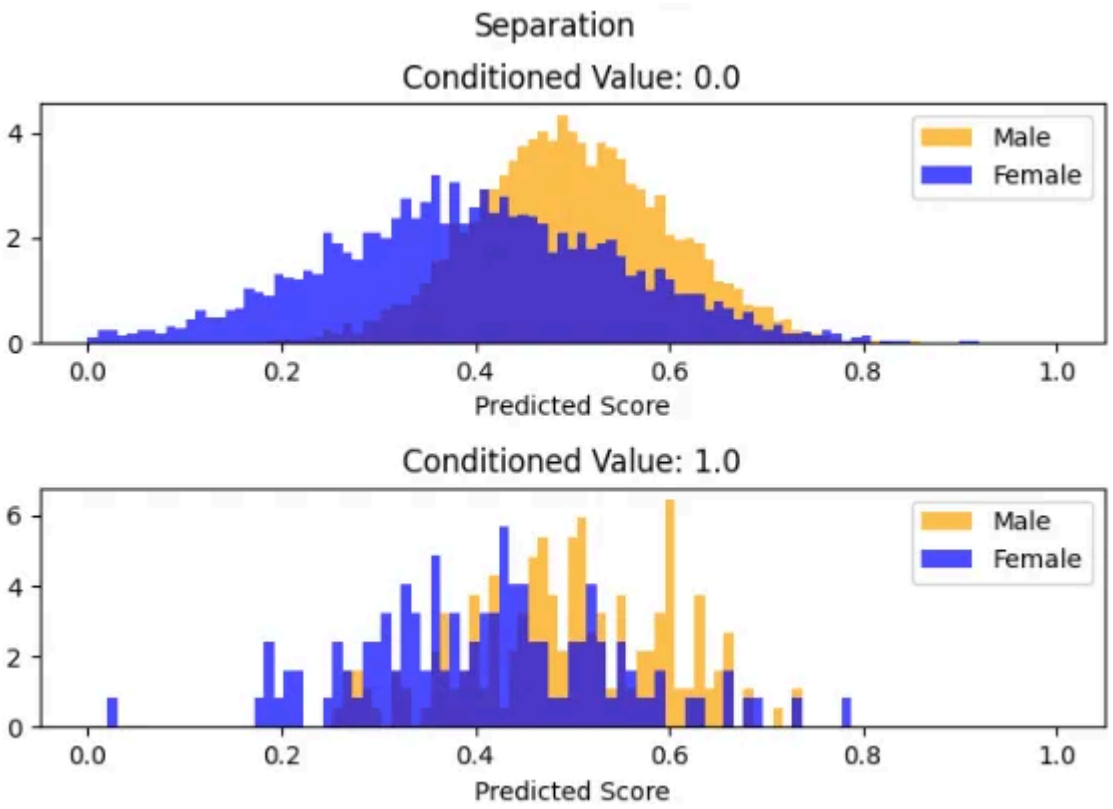
These probabilities depict that the sufficiency criterion was not met. These numbers lead us to the conclusion that the model isn't as accurate as predicting the appropriate qualifications of males as compared to females, implying that the model is biased towards males. Rectifying this is in the interest of the college as it accepts more qualified students.

Practical Visualization

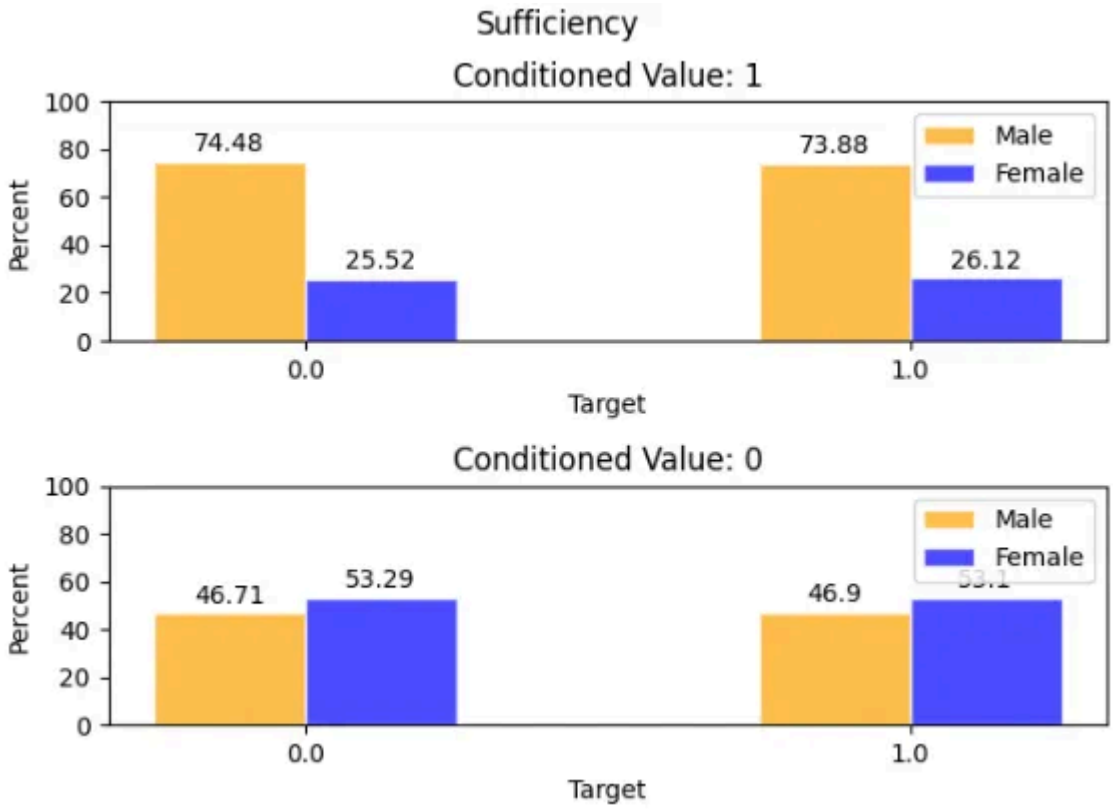
In order to drive these concepts home, I manually created some graphs to elucidate how these criterion will look visually in practice. Once again, let us assume these results came from a binary classifier in which the protected attribute is gender having values male and female. In most scenarios, the output of machine learning models return scores indicates its preference for the positive class i.e we don't have the predicted LABELS as shown in the above examples but rather the predicted SCORES. We want to ensure that the distributions of these predicted scores for each value of the protected attribute are almost indistinguishable based on the definitions of the aforementioned criterion. In order to decide whether or not the distributions are similar enough or not, you can choose a statistical test based on the properties you want satisfied.



For this model to satisfy the independence criteria, both the male and female distributions of the predicted scores should be very similar.



Separation requires us to condition on each target label and since we have two in our example, we get two different graphs. The male and female distributions within each conditioned value should be very similar in order to satisfy separation.



Visualizing sufficiency is a little more complicated because it is conditioned on the predicted scores which is a continuous random variable. In order to create this graph, I binned the scores into two categories 0 and 1, with 0 referring to the lower 50% scores and 1 referring to higher 50% scores (quantile binning). Interpreting this graph is pretty simple, it shows the percentage of males and females that have a true value of False and True when the predicted scores has a binned value of 0 (low) and 1 (high) respectively.

In each of these graphs, it is quite easy to observe that this model is biased towards males.

Impossibility Theorem of Fairness

How do I decide which criterion to select? Can I satisfy more than one? These are some of the questions that jump to mind after understanding these three criteria. The first thing to realize is that only one out of three criteria can be satisfied, unless it is a case where there isn't any correlation between the sensitive attribute and the target variable (also known as a degenerate/redundant case). I won't include the proof of this article because I don't consider it important in understanding how to measure bias in models. However, given this result, it becomes really important that the data scientist select the correct criteria to satisfy based on the problem they are trying to solve.

What Next?

Great! Now, that you have a basic understanding of measuring bias in machine learning models, you might wonder how to further your progress in this area? I've compiled a list of resources, albeit rather random, that should assist you in ensuring the fairness of your machine learning models!

- A Berkeley course on fairness in machine learning
- According to me, the best book out there for an introduction into this topic, Fairness and Machine Learning
- Equality of Opportunity in Supervised Learning, a research paper going in-depth into each of the criteria
- AIF360, an API to detect and mitigate bias, package available in both Python and R
- Fairlearn, another Python package for bias evaluation

I am extremely grateful to Geoffrey Schneider, my mentor at Vanguard for his valuable feedback and Belinda Li, for drawing the comic at the beginning of the article.

Machine Learning | Fairness | Artificial Intelligence | Bias | Ethical Ai

Follow

Written by Rahul Shekhar

49 Followers · Writer for Analytics Vidhya

Penn' 20

More from Rahul Shekhar and Analytics Vidhya

 Kia Eisinga in Analytics Vidhya