Accuracy

3 min

Great! We collected our measurements, made decisions about handling missing data. Now we need to ask ourselves if the dataset we have really describes the world. We need to know if it is accurate. Accuracy is a measure of how well records reflect reality.

While doing some Exploratory Data Analysis you notice that the trees you measured are overall taller than the trees I measured. That's interesting. You're not really sure why that is, so we compare how we measured the trees.

We realize that you measured starting from the ground and I measured starting from where the roots become the trunk. It's not a huge difference, but it's enough to affect the accuracy of our data. The tree heights are not accurate because we don't know how tall each tree really is. We could also say that the height variable is not *reliable*. Without a standard measurement unit and standard method, comparing trees, or even getting an average tree height is impossible.

Standardization is essential for accuracy – but it's not the only way that accuracy can be compromised.

There are a lot of ways a dataset can have low accuracy, but it all comes down to the question of: "are these measurements (or categorizations) correct?" It requires a critical evaluation of your specific dataset to identify what the issues are, but there are a few ways to think about it.

- First, thinking about the data against expectations and common sense is crucial for spotting issues with accuracy. You can do this by inspecting the distribution and outliers to get clues about what the data looks like.
- Second, critically considering how error could have crept in during the data collection process will help you group and evaluate the data to uncover systematic inconsistencies.
- Finally, identifying ways that duplicate values could have been created goes a long way towards ensuring that reality is only represented once in your data. A useful technique is to distinguish between what was human collected versus programmatically generated and using that distinction to segment the data.

Holding these perspectives in mind is important for both numeric and categorical variables. In fact, they often provide clues about each other.

As far as resolving accuracy issues, there's no simple solutions, and every solution has to be tailored to that specific dataset. In the end, the only way to improve a dataset's accuracy is to use real-world knowledge to be sure that the dataset reflects reality. But even then, the reason that we collect data is generally to learn something new about the world. Sometimes the data will surprise you, but distinguishing between a new finding and inaccuracy is the work of a skilled data scientist.

Instructions

Notice in the image that the measurements begin at different places. How else could inconsistencies be created during the data collection phase?