

Dewey Defeats Truman: How Sampling Bias can Ruin Your Model



ODSC - Open Data Science

Follow

Dec 1, 2018 · 4 min read



In the 1948 election season, Thomas Dewey faced off against incumbent Harry Truman for the presidency, running on the Republican and Democratic ticket, respectively. The Chicago Daily Tribune, a Republican-leaning paper at the time, ran a poll forecasting the outcome of the election, with a decisive win for Dewey on November 6th. The night before the election was called, the Tribune went to press with the now-infamous headline: DEWEY DEFEATS TRUMAN.

Of course, that isn't what happened. President Truman was elected with a comfortable margin of 49.6 percent of the popular vote to Dewey's 45.1 percent. In the Electoral College, Truman won 303–189. At the end of the day, the Tribune's polling predictions were off dramatically.

When the Tribune revisited their poll to see what went wrong, they quickly discovered that they had oversampled Republicans in their data for a pretty simple reason: the poll was conducted entirely over the phone. Since wealthy people were more likely to have a

phone and were also more likely to identify as Republican, the poll was skewed significantly towards Dewey.

This is one of the most famous examples of *sampling bias*, a phenomenon in which the data sample is selected in such a way that it fails to reflect the true underlying distribution.

To simplify the problem, say you have a basket of apples and oranges and want to figure out what percentage of the basket is made up of each. If you base your guess off of a sample where you intentionally picked out apples, that's not going to tell you very much about what the rest of the basket actually looks like.

What about modern sampling bias?

In **machine learning**, this is a recipe for disaster. Unlike humans, your model needs examples of every class you're trying to identify; it's not good enough to just have things that *aren't* your class and to extrapolate from there, even if you're dealing with a binary classification problem (this is probably obvious to any professional, but it might be surprising for some rookies).

For that reason, it's important to understand a few things about your data before you go about building your model. Chief among those considerations is whether there are enough examples of each class in order to solve the problem you're tackling.

Even more importantly, you need to consider whether the data sample you're using actually reflects the same underlying distribution you're attempting to model. However, this is usually an impossible task since we don't have access to the underlying function. If we did, we wouldn't need machine learning in the first place, because we could just use the original distribution to make predictions.

We still have an antidote to this problem though: *random sampling with replacement*. Because we're selecting our data sample without cherry-picking examples, we should, in theory, wind up with a data sample that reflects the true underlying distribution with relatively little bias.

Performing this operation with replacement means that selecting an example does not exclude you from selecting the same example as you continue to sample your dataset.

It's as if you reached into your basket, picked an apple or orange, wrote down the result and put it back before shaking up the bin and repeating.

Sampling with replacement is crucial to machine learning problems since excluding examples you've already seen will skew your distribution further and further each time you attempt to sample data. With our basket, say you remove 100 out of 200 pieces of fruit in the basket, piece by piece. Each time you take a piece out, the basket looks less and less like the original. By the time you get to 100 pieces, it could very well be useless at demonstrating what the basket is supposed to look like. However, if we put back the fruit each time, we preserve the original data sample, making sure that the distribution is intact each time we reach in.

How would things be different?



Let's look at Dewey and Truman one more time to make sure we understand the problem we've described. The Chicago Times Tribune introduced bias into their model by selecting too many people from a specific group of voters (wealthy people) and not enough of their second class of voters.

Another point worth noting: we don't know whether they called the same person twice, but if they did, it's entirely possible that the voter would reply, "but I already told you how I voted!" In a random sample, that shouldn't matter — those people need to be recorded twice in order to truly preserve randomness in the data sample.

Sampling bias will ruin your model, and you'll pay the cost for it if you decide to train on a biased data set. Make sure you don't fall into this trap when you're working on your next project.

Original story here.

Read more data science articles on [OpenDataScience.com](https://opendata-science.com), including tutorials and guides from beginner to advanced levels! Subscribe to our weekly newsletter here and receive the latest news every Thursday.

[Politics](#)[Data Science](#)[Sampling Bias](#)[Artificial Intelligence](#)[Technology](#)[About](#) [Help](#) [Legal](#)[Get the Medium app](#) A button that says 'Download on the App Store', and if clicked it will lead you to the iOS App store A button that says 'Get it on, Google Play', and if clicked it will lead you to the Google Play store