✦ Member-only story

Don’t Forget Confidence Intervals for Your ML Product

Machine Learning is never 100% correct. Thus, an ML model is only helpful when users understand the uncertainty of predictions.

Benjamin Thürer · Follow

Published in Towards Data Science · 7 min read · Oct 10

Almost every day we discover the launch of a new machine-learning product, service, or dataset. It is the age of AI and, yet, rarely do any of these products inform on how much confidence the user should have in the results. However, as research shows, good decision-making requires knowledge of when to trust AI and when not. Otherwise, it leads to the common situation that users need to try out the model frequently to get an understanding of when to trust and when not to trust that model and to find out if the offered product is useful for them.

The reason for such a try-and-error principle by the user is that every model (doesn’t matter whether it is based on ML or statistics) is built on data and its uncertainty. The model’s underlying data does not represent the actual ground truth of what the model is supposed to predict. Otherwise, if that ground truth would be available, you would not need a model in the first place. Thus, the resulting model will only provide an estimate and not a truth value.

> Put short, the correctness of machine learning and statistial models is uncertain and cannot always be trusted.

Example: Predict Cross County Movements

Let's take an example (Figure 1). Imagine a product that provides you with the amount of people moving from one country to another. Of course, there exists data (like tax reports) that will provide that information but would that data truly represent the full population of movements? Is every student, immigrant, or ex-pad changing their tax report? No, that is most likely not the case. Hence, even a straightforward product such as providing movements is biased towards its underlying data sample (for instance publicly available tax reports). It is easy to imagine how more sophisticated products can be biased.

For machine learning, this limitation gets even worse simply because of its probabilistic nature, the manifold input, and each input representing only a tiny part of the population. Thus, the underlying model will be biased towards the majority of cases described in the training data and will diverge from the underlying real-world situation without us knowing. Put short, the correctness of machine learning and statistical models is uncertain and cannot always be trusted.

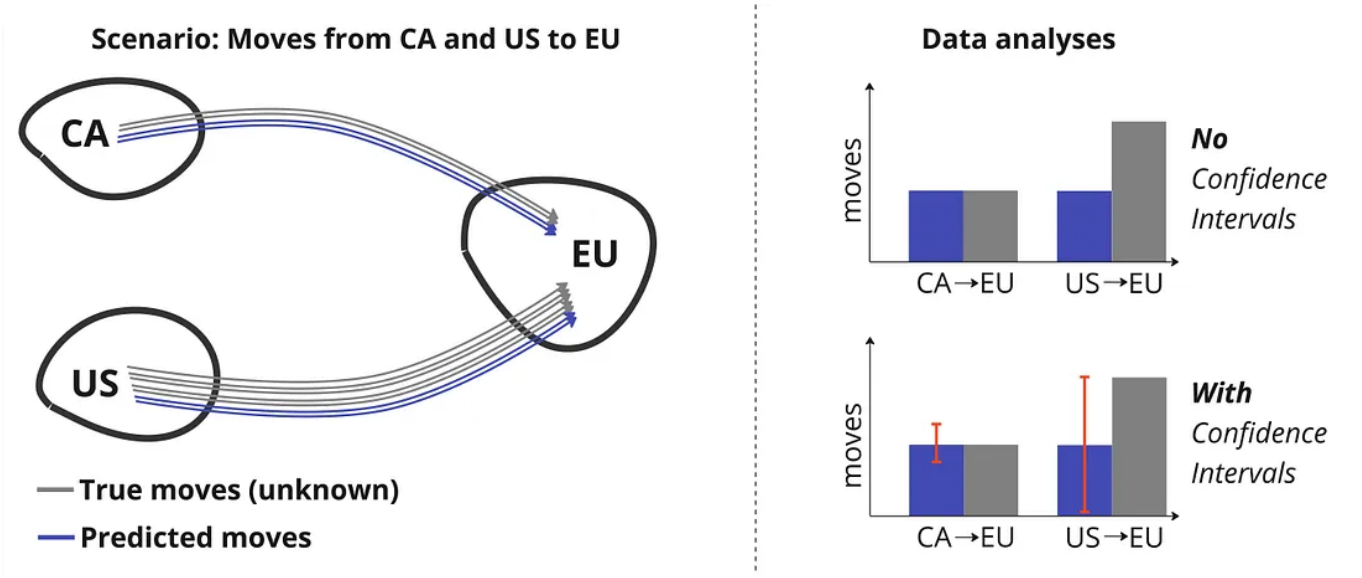

Figure 1: Example of a model predicting moves from one country to another. The true movements (gray) are not known to the user. Without understanding the confidence of the model, the interpretation of results might

be misleading. In this case, assuming that the same amount of people moved to the EU from CA as from the US.

Confidence Intervals Leave Decisions to the User

The reason why we sometimes *blindly* trust data products is due to our trust in the people or the company developing the product. We do expect a company to rigorously test a product for quality before launching it. But in a capitalistic world where more and more work is also outsourced to AI, can we trust that every product launch is good enough and properly tested?

As a matter of fact, **no** data product (and especially no machine learning model) will tell you when it is wrong. It will just give you a prediction (no matter how much that is off). So who is to blame if an important business decision is being made on wrong predictions? As a data scientist developing that product, I don't want that to be me! That is why providing intuitive confidence intervals is important for every data product.

> ***Confidence intervals are the solution that informs customers about the uncertainty of the product so they can make an educated decision whether to trust the predictions or not.***

That sounds complicated at first, and some users might be scared just of the term *confidence interval,* but they are not as scary as they sound like. Intuitive confidence intervals are helpful and are a sign of product quality. They show that the company cares for you as a user because they are trying to help you make the best decision.

What is a 95% Confidence Interval?

Choosing a 95% confidence interval is a common approach to describe confidence due to its ease of interpretation. The confidence interval comes with a lower and upper bound of the data (mostly displayed as slim bars on top of the chart). The lower and upper bounds basically describe a range or a corridor in which 95% of the predictions would fall given the actual (unknown) true value. Hence, a large range (compared to the predicted

value) would indicate that the user should have less confidence in the presented value itself since the underlying real value might be way more of.

How to Calculate Confidence Intervals?

Let's take the example from above and define a product that provides the amount of moves from Canada and the US towards EU countries. As seen in Figure 1, the hypothetical model here predicts moves closer to the ground truth for outflows from Canada compared to the US. Therefore, it would be beneficial to inform the end user about a high confidence in the Canada model compared to a lower confidence in the US model.

If we do not provide this information in the form of confidence intervals, the intuitive conclusion of that data (blue bars on the right since you would not know the gray bars) is that there is an equal amount of outflows between those two countries into the EU. However, as soon as we add confidence intervals describing how much trust we should put into those numbers the interpretation changes. If the confidence intervals would describe 95% confidence, a better interpretation is that the real underlying amount of moves might differ substantially from what is provided here. It could easily be twice as much and would still be in a range of 95% confidence. Hence, the user should not make a business decision for those moves on the data presented alone and rather focus on the CA moves or evaluate other data sources.

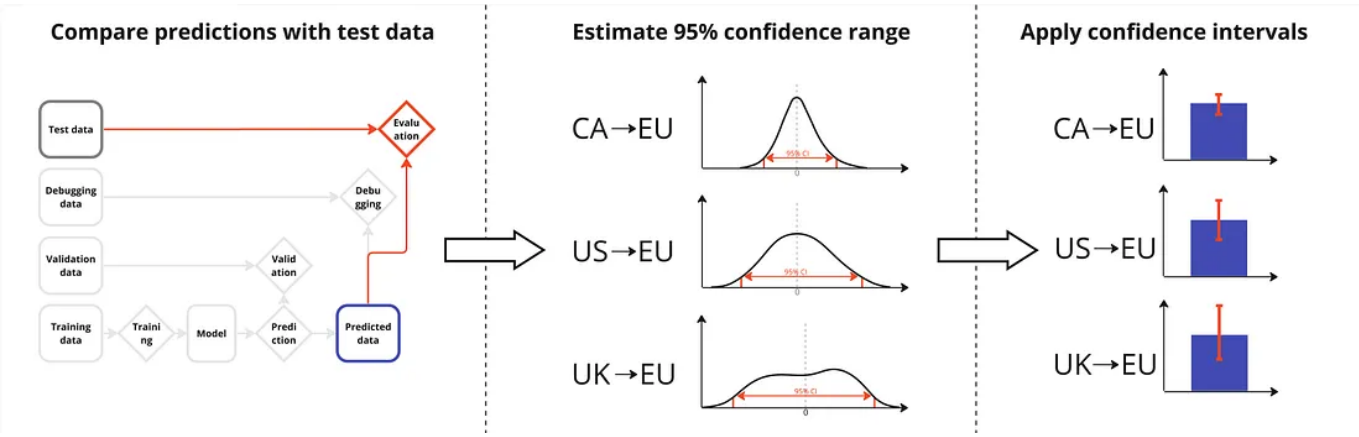

Figure 2: When your model is positively evaluated with your test data and is put into production, you can use that test data to calculate the difference between true and predicted values which, in turn, allows you to estimate a confidence range.

So how can we create and add confidence intervals to our models? As the name implies, to build confidence intervals (or any other measure of confidence) you first need to define what your metric represents and what confidence means to you and your user. There are many ways to derive a measure of confidence but all depend on some sort of knowledge which you can use to validate and test your predictions. For instance:

- **When you have ground truth data:** often ground truth data is available historically or coming in after a certain lag. If so, that can be used to inform how far "off" your predictions were and you can use that to update your confidence for future predictions. For instance, when you have access to tax reports indicating how many people moved you can see how well your predictions correlate with that.
- **When you have expert knowledge about your predictions:** For a lot of products, ground truth is not always available or comes with a very long delay. However, sometimes you can define certain traits or patterns your data should follow. If so, you could see those as correlates to ground truth and let them help you build confidence intervals. For instance, when you have expert knowledge of how movements should have been impacted by Covid19 you can analyze how well your predictions match that knowledge.
- **When you have a test set evaluation:** Even without ongoing ground truth or expert knowledge you are still able to describe confidence intervals using your test set evaluation (see Figure 2). Every model that is built requires a test set to ensure that the model is successfully predicting the desired outcome (see here for more info on model development). That test set will tell you also where your model performs well and where not. At least for in-sample categories, it will allow you to provide confidence intervals. The disadvantage here is, though, that models are prone to drift over time. So it is important to find a way to update your confidence intervals.

Simple Code Example (SQL)

When you have one of the above, Estimating a 95% confidence interval can be fairly straightforward as long as the data is normally distributed and each sample is independent of another (no autocorrelation). Here is an example of how confidence intervals for each category (primary key) could be estimated in standard SQL.

First, we calculate the error between our predictions and the ground truth:

```
SELECT
  date,
  primary_key,
  (prediction - groundtruth) AS error
FROM input_table
```

Then, a margin of error based on the standard deviation and sample size is calculated:

```
WITH base as (
SELECT
  primary_key,
  count(*) AS n,
  STDDEV(error) AS error_std
FROM error_table
GROUP BY primary_key
)

-- 1.96 represents z-score for 95% confidence level
SELECT
  primary_key,
  1.96 * SAFE_DIVIDE(error_std, sqrt(n)) AS margin_of_error
```

Finally, apply the margin of error to derive lower and upper confidence intervals:

```
SELECT
  date,
  primary_key,
  prediction - margin_of_error as lower_confidence_interval,
```

```
  prediction + margin_of_error as upper_confidence_interval
FROM input_table
JOIN margin_of_error_table USING(primary_key)
```

Summarised

Statistical and machine learning models provide estimates, not ground truth. We all must understand that and find our way of working with that. Measure of confidence will allow the user to understand how much trust and weight to put in the answers they receive from a certain model-based product and that will allow the user to make better decisions.
It is important to highlight that a low confidence score does not necessarily mean the product is of low quality, it might just mean that the product is not designed for the use case in hand and that other data sources should have more weight in the decision-making process.

It can be concluded that confidence intervals provide a helping hand to the end-user to make informed decisions and it also protects the data scientist from a too-simple interpretation that every data is always correct.

All images, unless otherwise noted, are by the author.

Machine Learning | ML Engineering | Data Science | Decision Making | Editors Pick