Probabilities and Percentiles: Part 3 - How They're Different, and Why it Matters

Machine learning is everywhere these days. One of the challenges we run into is understanding how to use the outcomes from the models we create. This problem is amplified by the fact that the terminology data scientists use is often unfamiliar to people without a strong mathematical background.

*Two of the most common ways to talk about the predictions from a machine learning model are the *probability* predicted by the model and the *percentile* of the prediction across the population evaluated by the model. It gets even more confusing since probabilities are often expressed as percentages. In this blog series, we will discuss what each of these outputs tell you and how to use them to solve problems for your business.*

This is the third of three blog posts, where we highlight the differences between probabilities and percentiles, and discuss when each output is most useful. For an introduction to probabilities, check out Part 1. For an introduction to quantiles, including deciles and percentiles, read Part 2.

Probabilities and Quantiles: How They're Different

To quickly recap our previous two posts in the series, the probabilities generated by a predictive model can be useful for understanding how likely an individual record is to belong to one outcome or another (e.g., person x has a 0.2 percent probability of buying product y). However, without knowing what the overall population looks like they are not helpful for making statements on where an individual record falls relative to the population of records.

Quantiles like deciles and percentiles are helpful for contextualizing an individual probability within the population of probabilities that are likely given a certain problem. They are the ranked groups a population can be divided up into. Knowing something is in the 20th percentile means that 20% of values in the population are less than that value, and 80% are greater. It doesn't necessarily say anything about the individual's probability, but it lets us compare it to the population at large.

When to us Probabilities or Percentiles

Both probabilities and quantiles (e.g., percentiles) can be useful for creating an actionable insight. It all depends on what your business goal might be.

Say, for example, we would like to increase the profitability of our sales funnel.

If we know the **cost** of pursuing a lead and the potential **payoff** of a converted lead, then we can use a ***probability*** to identify leads that are likely to be profitable and leads that aren't.

This approach can take into account multiple factors, including not only the probability that a given lead will convert, but also the potential lifetime value of a lead and the costs associated with different approaches. We are able to identify the most valuable leads as a function of both probability to covert and possible payoff, and compare that to cost to ensure we spend money pursuing leads wisely.

In the case for probabilities, we are looking for return on investment. By focusing on the highest possible value leads, we can bring in more money using the same investment.

In cases where we do not have as much of an understanding of the costs or potential payoffs of a lead, we can use the ***quantiles*** to only target the **best** leads (e.g. by targeting those in deciles 3-10, excluding the bottom 20% of leads).

In the case for quantiles, we are looking at improving our conversion rate by dropping the bottom few deciles from consideration. Here we are reducing out costs by reducing the overall number of leads we pursue, dropping those that are least likely to convert.

These examples illustrate the conditions when each metric is more beneficial:

- we like to use **probability** when we can estimate costs and payoffs of good and bad outcomes
- we like to use **quantiles** when we can't

This makes quantiles much more versatile, and it's the reason customers get back a decile score with any model Fenris provides.

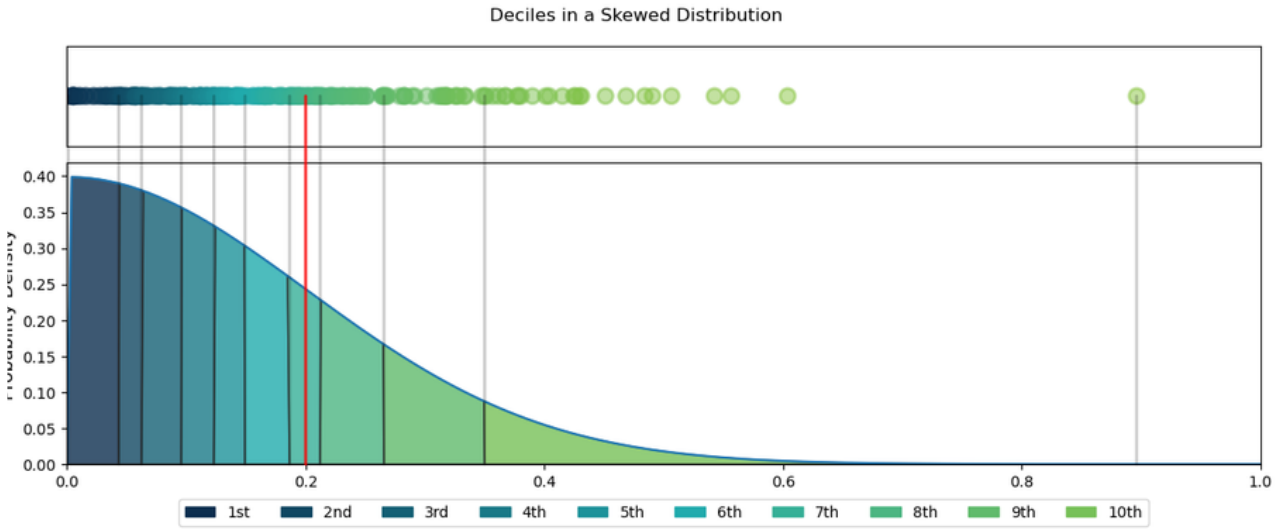
Why The Difference Matters

Imagine we had a goal where we wanted to improve our conversion rate by focusing our efforts on the leads that are most likely to buy, and dropping out the leads that are least likely to buy.

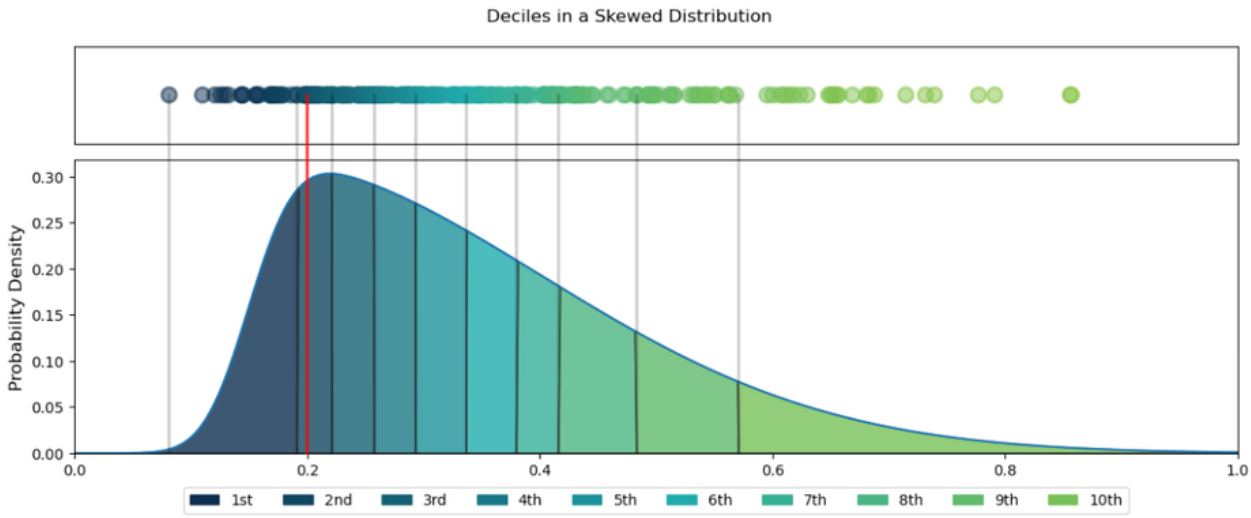
If we have person who has a 20% probability of buying our product, we might think this isn't a very good person to pursue. However, with deciles we might see that this person is actually in the 7th decile of leads, meaning that they are in the top 40% of people likely to buy our product.

However, if we have a product that is a more popular seller, a person with a 20% percent probability might be in the second decile, meaning that they are in the bottom 20% of people likely to buy our product.

This is something we can visualize in the following two figures. The first figure shows an even with a 20% probability that falls into the 7th decile.



And the second figure shows a 20% probability falling into the 2nd decile.



In this example we have seen two events with the same probability belong to very different quantiles because of what the overall distribution of probabilities generated by the model looks like.

If we wanted to improve our overall conversion rate by dropping the bottom 20% of leads, in our first distribution our 20% probability to buy lead would be included, but in the second distribution, the 20% probability lead would be excluded.

This is why quantiles are so useful - they contextualize the probabilities we return within the larger population and problem space. Treating probabilities like percentiles might result in you throwing away all of your best opportunities, or not discarding the proportion of lower probability items you were hoping to discard.

Talk to Us

This blog series is a result of all of our experience working with our clients. Do you have a similar use case? Or maybe something entirely different? We would love to work with you to learn about your problem, and how we can help.

Resources

Plots in this blog series were modified from https://github.com/mGalarnyk/Python_Tutorials/blob/master/Statistics/boxplot/box_plot.ipynb