BUZZ CHRONICLES > MACHINE LEARNING
Saved by @SteveeRogerr
See On Twitter

Twitter Thread by Cassie Kozyrkov

Cassie Kozyrkov
@quaesita

1/■ Good #DataScience advice that breaks pretty much every rule you learned in class... a thread. (+full blog post linked)

English version: https://t.co/dG4l6vPFBT
Spanish version: https://t.co/gFAjPQ5clS

#AI #MachineLearning #Statistics #RStats

2/■ Allow your approach to be sloppy at first and burn some of your initial time, energy, and data on informing a good direction later. That's right, you're supposed to start sloppily ON PURPOSE.

3/■ Have a phase where the only result you’re after is *an idea of how to design your ultimate approach better.*

4/■ In other words, start with a pilot phase where the objective isn't finding answers, it's finding a good approach to finding answers.

5/■ That means you're encouraged (ENCOURAGED!) to start with everything your stats classes told you not to do:

6/■ Low-quality data: use small sample sizes, synthetic data, and non-randomly sampled data to gain insights about the data collection process itself.

7/■ Rough-and-dirty models: seek an understanding of what the payoff from minimum effort looks like. Start with bad algorithms which you know are only going to give you a benchmark, not your best solution.

8/■ Multiple comparisons: instead of picking a single hypothesis test, feel free to throw the kitchen sink at your data to discover signals worth basing your final approach on. Add deadlines and MVP milestones to avoid the trap of infinite polishing, poking, and prodding.

9/■ If the statistician in you isn’t screaming yet, I admire your sangfroid. This advice breaks pretty much every rule you learned in class. So why am I endorsing these “bad behaviors”?

10/■ So why am I endorsing these “bad behaviors”? Because this is the pilot phase. I’m all about following the standard advice later, but this early phase has different rules.

11/■ The important thing is to avoid rookie mistakes by remembering these 2 crucial principles:

12/■ Principle 1: Don’t take any findings from the early phase too seriously.

13/■ Principle 2: Always collect a clean new dataset when you’re ready for the final version.

For more info: https://t.co/Ue332SMjy1

14/■ You’re using your initial iterative exploratory efforts to inform your eventual approach (which you’ll take just as seriously as the most studious statistician would). The trick is to use the best of exploratory nimbleness to inform what’s worth considering along the way.

15/■ If you’re used to the rigidity of traditional statistical inference, it’s time to rediscover the benefits of pilot studies in science and find ways to embed the equivalent into your data science projects.

16/■ The key thing to understand about this advice is that

- finding good questions
- finding good answers
- finding good approaches going from one to the other

are all different objectives that require different approaches. Sometimes there's homework to do before answers...