Twitter Thread by Mark Tenenholtz



Mark Tenenholtz
@marktenenholtz

The worst taught skill in machine learning is model validation.

If you can’t validate your models well, you have no idea if they will actually work.

Here are 3 steps I’d take if I was relearning model validation from scratch ■

1. Learn the essential evaluation metrics

Think accuracy should be your primary metric? You’re sorely mistaken.

Most of the best metrics instead focus on how far your were from the correct answer. Think RMSE and MAE.

Others point to how well calibrated your model is, like F1.

2. Learn the common forms of cross validation

Before diving in too deep, make sure you understand the basics.

You can’t become an expert in validation in the classroom, but knowing what is out there (simple k-fold, stratified, grouped, roll forward, etc.) is crucial.

3. Read old Kaggle competition solutions

Every day, or multiple times a week, pick an old Kaggle competition.

Read every solution that is posted and skip to their validation schemes.

There are nuances to every dataset, and this is the best way to see how pros navigate them.

4. Build simple models and try different CV schemes

Get a dataset and create a random test set.

Then, build some simple models and switch validation strategies in and out and see how well your models generalize for each scheme.

This will cement the importance of validation.

5. Go and do it. A lot.

You will only improve at validation if you apply it to a ton of datasets.

If you stop after step 2, your skills will not be good enough. Full stop.

Never rest on your laurels. There is always something new to learn, and some new trick you can use.

This is a pretty general outline, but I plan on diving into the specifics on evaluation metrics and CV schemes in the future.

I also discussed them on a podcast with @bhutanisanyam1 here: https://t.co/AiGAe1zBH3

Follow me @marktenenholtz so that you don't miss it!