QUANT RUSH · PICKS & SHOVELS · WORKSHOP 4

WHAT GLITTERS

Validation and overfitting: why the public leaderboard lies.

TUE NOV 10

NOT ALL THAT GLITTERS IS GOLD



Lithograph: Kellogg & Comstock, “California Gold Diggers,” c. 1850 · Yale Beinecke · Public domain

TODAY

Three ways a score glitters

1

LEAKS

Your check lets the model see answers the test set never gives.

2

OVERFITTING

The model memorizes its training eras instead of learning what repeats.

3

NOISE

A score on a few hundred eras is partly luck.

Then: the leakage checklist, your two finals, and this week's exercises.

THE PUBLIC BOARD

A score is gold only if it holds on eras the model never saw.

The public board scores about 30% of the test set. The other 70% decides the ranking, and you see it only after the deadline.

Fool’s Gold goes to the public top-10 team that falls furthest.

LEAKS

Random rows flatter a model

Two GBMs, the same 1,200 training eras, two ways to split them. Then the truth: train on all 1,200 and score the 1,080 later eras.

Scored by	Steady GBM	Flexible GBM
Random rows, 4 folds	0.1113	0.1129
Era blocks with a 50-era gap	0.1055	0.0966
Truth: later eras	0.0881	0.0826

Random rows rate the flexible GBM higher. The truth says it is worse.

VALIDATE LIKE THE FUTURE

Whole eras, in order, with a gap

1

KEEP ERAS WHOLE

The market columns are the same on every row of an era. A flexible model can recognize the era and reuse its answers.

2

PAST, THEN FUTURE

Train on earlier eras, validate on later ones. The test set comes after the training set.

3

LEAVE A GAP

Neighbouring eras share recent history. Skip 50 eras between the two sides.

At test time you never have a single answer from the eras you predict. Your validation shouldn't either.

WALK-FORWARD VALIDATION

Five honest checks, not one



FOLD	STEADY GBM	RIDGE	-ret_00
1	0.1236	0.1015	0.0839
2	0.1096	0.0949	0.0751
3	0.1047	0.0881	0.0673
4	0.1005	0.0769	0.0490
5	0.0822	0.0754	0.0663
MEAN	**0.1041**	**0.0873**	**0.0683**

Trained on earlier eras only, with a 50-era gap.

The GBM beats Ridge in all 5 folds, and scores mostly drift down with time.

OVERFITTING

Memorized is not learned

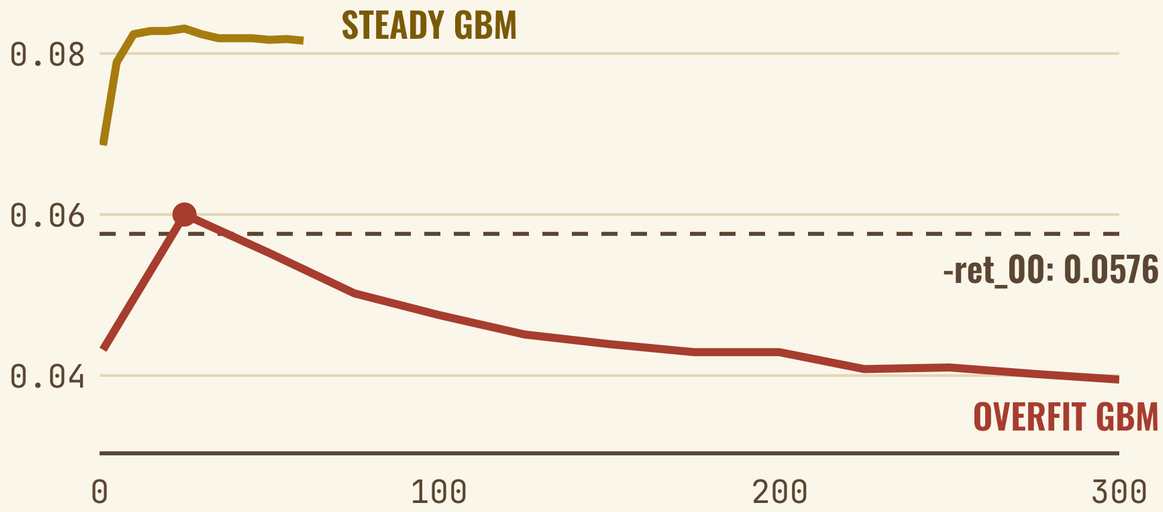
Each model trained on the same 500 eras, then scored on those eras and on the 1,080 recent eras.

Model	Its training eras	Recent eras
Ridge	0.0940	0.0733
Steady GBM: 15 leaves, 60 rounds	0.1859	0.0816
Overfit GBM: 255 leaves, 300 rounds	0.9542	0.0395

Every model looks better on its own training eras. The overfit one ends below the one-line rule, -ret_00 (0.0576).

OVERFITTING, ROUND BY ROUND

It peaks early, then falls



SCORE ON THE 1,080 RECENT ERAS, BY BOOSTING ROUND

Overfit GBM: peaks near round 25 (0.0600), then falls, down to 0.0395 at round 300.

Steady GBM: levels off near 0.082 and stays there.

The cures: smaller trees, bigger leaves, fewer rounds, more training eras, and settings chosen by validation.

NOISE

Close scores are ties

A is the flexible GBM from the split test; B is the steady GBM trained on 500 eras. Both are scored on the recent eras.

ALL 1,080 ERAS

A 0.0826 · B 0.0816

The gap, +0.0010, is smaller than its uncertainty (±0.0017). A wins in 50% of eras.

2,000 RANDOM SAMPLES OF 30%

B ahead in 35%

A alone scores anywhere from 0.0730 to 0.0918 in 90% of the samples.

On a few hundred eras, a gap of a few thousandths tells you almost nothing.

THE WINNER'S CURSE

The best-looking score is partly luck

Twenty tweaks of one model, equally good by design. On each random 30% sample of the recent eras, keep the tweak that scores best.

0.0788

The 20 tweaks on all 1,080 eras, on average (0.0769 to 0.0808).

0.0818

The best-looking tweak, on the sample that picked it.

0.0799

The same pick on all 1,080 eras: barely above average.

Keep changes that have a reason and hold in most folds, not the ones that moved the board.

BEFORE YOU TRUST A SCORE

The leakage checklist

1 SPLIT BY TIME
Whole eras, earlier to later, with a gap between.

2 NO TARGET INSIDE
No input is computed from the target. At test time there is none.

3 LEARN ON TRAINING ERAS
Feature choices, scaling and settings too, never on validation eras.

4 THE BOARD IS A CHECK
Never tune to the public score.

5 COMPETITION FILES ONLY
No outside data of any kind.

6 AGAINST THE RULES
Matching rows across eras, or using the test set's structure to work out answers.

Winners' code is checked, and rerun with only the competition files.

YOUR TWO FINALS

Pick by validation, not by the board

Final 1

YOUR BEST VALIDATED

The best by walk-forward folds, not the best public score.

Final 2

A DIFFERENT, STURDY BET

A simpler model or a blend. If one stumbles, the other may hold.

Pick none, and Kaggle uses your best public scores. More at the Final sprint.

THIS WEEK

Try it in the notebook

1 FIND THE LEAK

Rerun the split test without the market columns. How much of the random-row leak is left?

2 CLOSE THE GAP

Set the gap to 0 eras. Do the era blocks get more optimistic?

3 TUNE HONESTLY

Change one setting and compare fold by fold. Keep it only if it wins in most folds.

4 SUBMIT

Train on every era, submit, and judge it by your folds, not the board.

Get the notebook at smcqfec.com, then import it into your own Kaggle notebook.

RECAP

What glitters, and what holds

VALIDATE LIKE THE FUTURE

Whole eras, in time order, with a gap. Random rows flatter.

WALK FORWARD

Several honest scores; compare models fold by fold.

WATCH THE GAP

Training far above validation means overfitting.

SMALL GAPS ARE TIES

The best public score is partly luck.

NEXT · TUE NOV 17 **The Last Dig** (Final sprint): ensembling and your two finals.