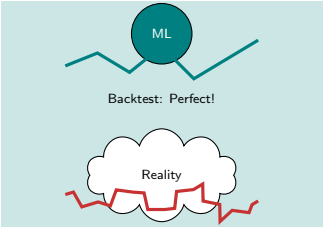
Why do most stock price prediction models fail even when they look brilliant in backtests?

The backtest trap:

- Models trained on past data memorize patterns that no longer exist
- Transaction costs and slippage invisible in simulations
- Overfitting creates false confidence in worthless signals
- Markets are non-stationary – what worked yesterday fails tomorrow

Why brilliant backtests fail in reality:

- Data snooping bias – testing hundreds of strategies and reporting the winner
- Look-ahead bias – using information unavailable at decision time
- Survivorship bias – excluding companies that went bankrupt
- Regime changes – crises break all historical patterns



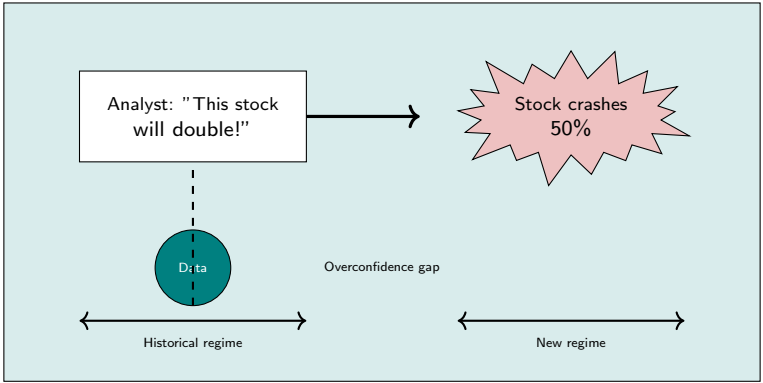
Live: Disaster

Core Insight

A backtest tells you what would have worked if the future resembled the past. Markets ensure the future never does.

The gap between backtest and live performance is where fortunes disappear.

Have you ever been certain about a prediction that turned out completely wrong?



Reflection

Overconfidence in predictions comes from confusing patterns in historical data with laws of nature. Markets have no laws – only temporary regularities that decay.

Certainty about future prices is inversely proportional to understanding how markets work.

What are the theoretical limits on predicting financial markets?

Three fundamental barriers:

- **Non-stationarity:** Statistical properties change over time – mean, variance, correlations all shift unpredictably
- **Adversarial environment:** Other traders compete away predictable patterns through arbitrage
- **Reflexivity:** Predictions influence the market, invalidating themselves

What the Efficient Market Hypothesis teaches:

- Weak form: prices reflect all past price data – technical analysis fails
- Semi-strong: prices reflect all public information – fundamental analysis has limited edge
- Markets are approximately efficient – alpha is scarce, temporary, and quickly competed away

Autocorrelation patterns:

- Return autocorrelation: within confidence bands
- Absolute return autocorrelation: highly significant
- Implication: predict risk, not reward

Regime shifts break models:

- Crisis regime: correlations spike to one
- Low volatility regime: patterns emerge
- Transition: unpredictable and catastrophic
- Models trained on one regime fail in the next

Core Insight

Returns have near-zero autocorrelation. Volatility is predictable; direction is not. This asymmetry defines what machine learning can and cannot achieve.

You can predict that volatility will cluster, but you cannot predict when the next regime change arrives.

How does overfitting turn a worthless model into a backtest champion?

The overfitting recipe:

- Start with hundreds of candidate features
- Test every combination on historical data
- Keep the model that maximizes backtest returns
- Ignore that you tested hundreds of worthless variants

Why it produces fake champions:

- With enough tries, random noise looks like signal
- Multiple testing without correction guarantees false discoveries
- Model memorizes specific historical accidents, not repeatable patterns
- Out-of-sample performance collapses because noise does not repeat

Core Insight

If you test one thousand trading rules at significance level five percent, fifty will pass by pure chance. Reporting the winner without multiple testing correction is data snooping.

Red flags of overfitting:

- Sharpe ratio above three
- Zero losing months in backtest
- Win rate above seventy percent
- Perfect fit to training data
- Hundreds of features for small dataset

In-sample vs out-of-sample:

- In-sample Sharpe: often above two
- Out-of-sample Sharpe: typically negative
- Transaction costs erase remaining edge
- Slippage and market impact finish it off

Overfitting is the primary reason most quantitative strategies fail in live trading.

How do legitimate forecasting and curve-fitting pipelines differ in structure?

Legitimate forecasting protocol:

- Walk-forward validation – train on past, test on strict future
- Never use future information in training
- Include transaction costs, slippage, market impact
- Report average out-of-sample performance, not cherry-picked best period
- Test on multiple time periods and market regimes

Curve-fitting warning signs:

- Random train-test splits on time series data
- Hyperparameter tuning on test set
- Survivorship bias – testing only on companies that survived
- Look-ahead bias – using restated data unavailable at decision time
- Reporting only the best-performing variant

Proper validation checklist:

- Temporal train-test split enforced
- Embargo period between train and test
- Transaction costs modeled realistically
- Slippage based on market depth
- Point-in-time data only
- Survivorship-free dataset
- Multiple testing correction applied

Realistic expectations:

- Sharpe ratio: between zero point eight and two
- Annual return: between eight and twenty-five percent
- Max drawdown: minus fifteen to minus thirty percent
- Win rate: fifty to fifty-five percent

Core Insight

Walk-forward validation is mandatory for time series. Random splits leak future information into training and guarantee overfitting.

The median quantitative hedge fund Sharpe ratio is zero point seven. Humility is a competitive advantage.

(c) Joerg Osterrieder 2025–2026

What happens when a prediction model starts influencing the market it is trying to predict?

The reflexivity problem:

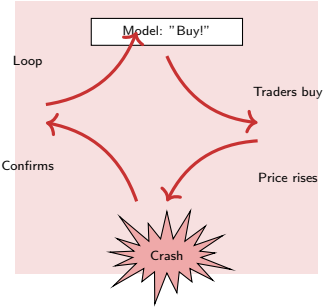
- Model predicts price will rise, so traders buy
- Buying pushes price up, confirming the prediction
- More traders adopt the model, amplifying the effect
- Eventually the feedback loop reverses catastrophically

Why this breaks prediction:

- The act of predicting changes the outcome
- Alpha decays as more participants follow the signal
- Crowded trades amplify volatility and drawdowns
- Published strategies stop working once widely known

Core Insight

Unlike physics, financial prediction is adversarial. Publishing a successful trading strategy guarantees it will stop working as others arbitrage it away.



Alpha decay timeline:

- Proprietary signal: weeks to months
- Published factor: three to seven years
- Viral social media signal: days

Alpha is not a stock – it is a decaying asset that vanishes when discovered.

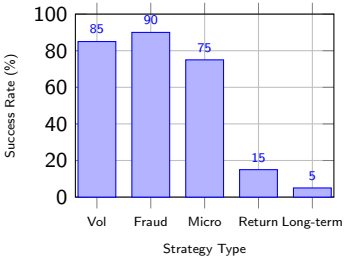
Where has quantitative prediction consistently succeeded and where has it consistently failed?

Success stories – what works:

- Volatility forecasting using historical patterns
- Anomaly detection for fraud and money laundering
- Market microstructure prediction at millisecond scale
- Risk management and portfolio optimization
- Sentiment extraction from text without trading on it directly

Failure patterns – what does not:

- Predicting stock price direction from past prices alone
- Long-term return forecasts beyond a few days
- Strategies that work in backtest but fail with real transaction costs
- Models trained in one market regime and deployed in another
- Any strategy promising Sharpe ratios above three



Pattern:

- High success: stationary targets
- Low success: non-stationary targets
- Volatility changes slowly
- Returns change constantly

Core Insight

Machine learning excels at predicting risk and detecting patterns in stable environments. It fails at predicting reward in adversarial, non-stationary markets.

Predicting volatility achieves eighty-five percent accuracy; predicting direction achieves fifty percent.

(c) Joerg Osterrieder 2025–2026

Who profits from selling prediction and who pays when predictions fail?

Who profits from selling predictions:

- Vendors selling backtested strategies with survivor bias
- Newsletter writers cherry-picking past calls
- Consultants charging for models they never trade themselves
- Platforms monetizing order flow from retail prediction believers

Who pays when predictions fail:

- Retail investors buying overfit strategies
- Institutional investors deploying models without proper validation
- Anyone confusing backtest performance with future returns
- Regulators cleaning up after algorithmic trading disasters

Core Insight

If a prediction strategy genuinely worked, the discoverer would trade it quietly, not sell it publicly. The act of selling reveals it no longer works.

The incentive problem:

- Genuine alpha is kept secret
- Failed strategies are repackaged and sold
- Marketing emphasizes backtest, hides live results
- Survivorship bias inflates apparent performance
- Fees extracted regardless of outcome

Case study – Long-Term Capital Management:

- Nobel Prize-winning models
- Leverage of twenty-five to one
- Four point six billion loss in nineteen ninety-eight
- Tail risk ignored
- Liquidity vanished when needed most

Follow the incentives: if the strategy worked, they would trade it, not sell it to you.

Three questions to distinguish genuine forecasting skill from statistical illusion

The Prediction Reality Test:

Question One: Does performance survive out-of-sample testing?

- Walk-forward validation on unseen data
- Multiple market regimes tested
- Transaction costs and slippage included
- Performance consistent across time periods

Question Two: Can you explain the economic reason behind the signal?

- Not just statistical correlation
- Behavioral or structural market inefficiency identified
- Mechanism linking inputs to outputs is clear
- Theory predicts when signal should fail

Question Three: Would the strategy still work if everyone followed it?

- Capacity constraints considered
- Arbitrage limits identified
- Alpha decay timeline estimated
- Reflexivity effects modeled

Application checklist:

- ✓ Out-of-sample Sharpe above one
- ✓ Economic story makes sense
- ✓ Works with realistic transaction costs
- ✓ Capacity limits understood
- ✓ Not widely published or crowded
- ✓ Survives multiple regimes
- ✓ Theory explains when it fails

Common failures:

- Perfect backtest, negative live performance
- No economic explanation offered
- Infinite capacity assumed
- Published years ago, still advertised as working
- Sharpe above three (implausible)

All three questions must pass. One failure disqualifies the entire strategy.

(c) Joerg Osterrieder 2025–2026

Your Challenge

Scenario:
You are presented with a backtest showing the following results over ten years of historical data:

- Annual return: sixty-two percent
- Sharpe ratio: three point four
- Maximum drawdown: minus four percent
- Win rate: seventy-eight percent
- Zero losing quarters

Task One: Identify three specific red flags that suggest overfitting.

Task Two: Propose one concrete test to distinguish real skill from luck.

Learning Goal

Healthy skepticism toward spectacular backtest results protects capital. The ability to identify overfitting is more valuable than any prediction algorithm.

If results look too good to be true, they almost certainly are.