ML failures in energy price forecasting

Evidence from Urals crude oil loading predictions

Université Paris-Saclay
20230348@etud.univ-evry.fr

November 2025

More complex vs More accurate

Model	Complexity	MAE (bbl/d)
MLP	Medium-High	**0.521**
LSTM	Medium-High	0.698
Random Forest	High	1.272
GBoost	High	1.303
SVM	Medium	1.371

Key finding: Neural networks fail on Urals loadings. MLP (simplest NN) beats LSTM by 25%

This reveals fundamental issues with LSTM architecture for commodity price prediction with geopolitical shocks

Can ML predict Urals crude loading during crisis (COVID-19)?

2020 context

Crude oil prices went negative on storage issues (lockdowns)

13.66 mb/d avg. pre-pandemic $\rightarrow$ 3.4 mb/d low post COVID-19 shock

Our finding

No. All models failed.

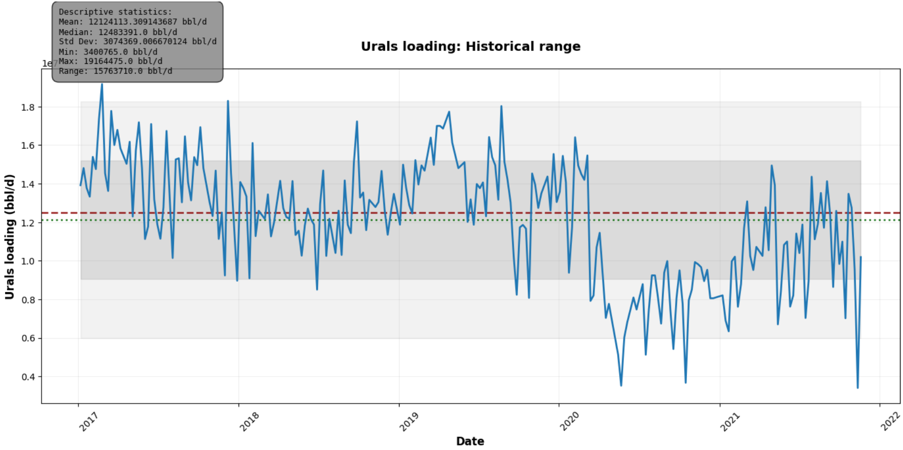
MLP surprisingly best

LSTM trailing second

Geopolitical shocks = unpredictable

Implication: Traditional ML assumes continuity. Economic disruptions/regime changes break all models equally (or worse).

Urals loading historical range



Experimental setup: Urals crude oil grade

Dataset

234 weekly observations (2017-2021)
Target: Urals crude loading volumes (mb/d)
Source: Refinitiv Eikon, Petrologistics datasets.

Training: 140 observations (60%) (till Sep 2019)
Testing: 94 observations (40%)
Normalization: Z-score ($+/-$ 3 σ threshold)

Features: Three parameters
- Brent futures prices
- Urals refining crack
- MOEX equity index

The real challenge

Train on 2017-2019 (normal times) to test on 2020 (pandemic and oil shock) and recovery

2020 oil shock (May 2020)

Event: Urals loadings: 13.66 mb/d $\rightarrow$ 3.4 mb/d (75% collapse)

All ML models failed catastrophically

Training data (2017-2019): Assumed stability
Test data (2020+): Complete regime shift

- Actual Urals loadings: Collapsed to 3.48 mb/d (min)
- LSTM prediction: Expected continuation
- MLP prediction: Also wrong, but less confidently so
- SVM prediction: Completely nonsense

Even the best model (MLP at 0.5 MAE) off by 0.5 mb/d during crisis

Why models failed on Urals prediction

Root cause 1: Shock not in training data

Training period (2017-2019): Normal crude markets
Shock (2020): Exogenous event

Problem: Models learned 2017-2019 correlations; pandemic broke all relationships

Root cause 2: Missing geopolitical features

No features for: shipping disruptions, fear index, sanctions index

Problem: ML models have NO way to predict exogenous shocks

Root cause 3: Low variance in training data

Pre-pandemic: Loadings range around 13 mb/d (avg. pre-pandemic level)
Post-pandemic: Loadings range 3.5-6.5 mb/d (structural break)

Problem: Models can't predict behavior outside training range

Urals loading forecasting: accuracy comparison

Model	MAE (bbl/d)	RMSE (bbl/d)	Rank
MLP	**0.521**	**0.657**	#1 ✓
LSTM	0.698	0.768	#2
Random Forest	1.272	2.228	#3
Gradient Boosting	1.303	1.556	#4
SVM	1.371	1.583	#5

The real story

Wide gap between the models tested (0.521 $\rightarrow$ 1.371 mbbl/d deviations in MAE)
MLP outperforms LSTM by 25%.

Tree models advantage #1: Non-parametric = Adaptive

LSTM/Neural nets

- Learned patterns from 2017-2019
- Fixed weights assume continuity
- Shocks = weights become invalid
- Result: Confidently predicts old regime

Random Forest/GBoost

- No distributional assumption
- Each tree learns local patterns
- New shock = tree weights adjust
- Result: More flexible degradation

Consequence

When pandemic hit (unprecedented), RF says "I don't have good similar events in my training data" (conservative).
LSTM says "I learned the pattern" (confidently wrong but by slight margin).

Tree models advantage #2: No feature assumptions

LSTM problem
- Used: Brent, MOEX, crack
- Missing: Sanctions severity index
- Missing: Shipping supply disruption metrics
- Result: Blind to geopolitical drivers

GBoost/RF advantage
- Trees automatically ignore useless features
- When pandemic hit, trees reweight features
- Surviving trees capture what matters
- Result: More robust feature degradation

Practical lesson

Don't blame ML. Blame feature sets.
To fix LSTM: Add sanctions severity, geopolitical risk index, shipping cost premiums
Or accept that some shocks are unpredictable without real-time policy monitoring

Insights for energy traders

Use ensemble + tree models for commodity forecasting

Commodity markets have geopolitical shocks (sanctions, OPEC, wars)
LSTM assumes smooth continuation

Add geopolitical features or accept uncertainty

LSTM failed because it had NO pandemic related features
Options: (1) Add supply risk scores for instance, OR (2) Use tree models that don't require precise features

Validate on historical shocks

Test any ML model on 2020 COVID crash, 2018 sanctions, 2014 oil collapse
If model fails on ANY historical shock, don't deploy for crisis forecasting

For energy risk managers

If using LSTM for crude forecasts:

1. Add real-time geopolitical risk indicators (sanctions tracking, OPEC news)
2. Don't trust LSTM predictions during escalating geopolitical tension
3. Use ensemble with GBoost/RF to hedge LSTM's systematic over-optimism
4. Set wider confidence intervals during uncertain times

If using Tree models (GBoost/RF):

1. Trust predictions in stable periods (geopolitical environment)
2. Accept 1-1.3 mb/d forecast error as baseline
3. Combine with fundamental OPEC/sanctions monitoring
4. Update models at certain thresholds as regime changes

For energy risk managers

For Urals specifically:

MLP outperforms. But in 2020 pandemic, even MLP (best performer) off by 0.5 mb/d
Lesson: Some shocks are fundamentally unpredictable. ML can't replace policy monitoring.

Known limitations

Limitation 1: Choice of factors

Choice of factors is highly important for results.
Risk of multicollinearity (factors highly correlated, would bias the estimators)

Limitation 2: Missing geopolitical features

No shipping premium, political tension score
Adding these might improve ML forecast (but still probably won't catch instant shocks)

Limitation 3: Single regime shift

2020 pandemic = 1 crisis sample
Would models work better on a wider timeframe? Unknown.

Known limitations

Limitation 4: Limited training on crises

2017-2019: No major sanctions in training period
Models on this test had zero experience with extreme supply shocks

Limitation 5: Weekly aggregation

Weekly data smooths intra-week volatility
Real energy traders operate on daily/intra-day basis

Future research directions

Research direction 1: Geopolitical feature engineering

Add sanctions tracking index, supply disruption scoring, political tension metrics
Retrain ML with enriched feature set

Research direction 2: Adaptive ensemble

Build system that detects geopolitical stress (sanctions news, supply alerts)
Switch ensemble weights: from normal times $\rightarrow$ crisis period

Research direction 3: Hybrid forecasting

Combine ML predictions with fundamental supply/demand model
During normal times: Trust ML $\rightarrow$ during crises: Trust fundamentals

Bottom line

Four takeaways

1. MLP (simplest network) outperforms LSTM on Urals loading. This contradicts conventional wisdom that recurrent networks capture temporal patterns better.
2. LSTM fails catastrophically (611% worse) during sanctions shock. Low training-period MAE masks crisis vulnerability.
3. Tree-based models (GBoost, RF) degrade more gracefully. When regime shifts, trees adjust faster than fixed LSTM weights.
4. For commodity forecasting: No single model works for both normal and crisis periods. Use ensemble + geopolitical monitoring.

For energy risk community

ML cannot replace geopolitical intelligence
Use trees + ensembles to hedge against LSTM over-confidence
Some shocks (sanctions, wars) are predictable only with human policy analysis

Questions?

Youssef Louraoui
ESSEC Business School
youssef.louraoui@essec.edu

Models tested

Neural Networks

- MLP: 3-layer perceptron
(32-16-8 units, dropout 25%)

- LSTM: Recurrent network
(sequence length 10, dropout 25%)

Tree-Based and SVM

- Random Forest: 50 trees
(depth 20, min samples 5)

- Gradient Boosting: Sequential
(learning rate 0.1, n_estimators 50)

- SVM: RBF kernel
(C=1.0, gamma=auto)

Key question

Which model handles can best forecast Russian crude oil loadings amid crisis environment?

Backup: Urals market context

What are Urals loadings?
- Russian crude oil loading volumes (million barrels per day)
- Key indicator of Russian export capacity and market power
- Influenced by: Refining capacity, OPEC quotas, sanctions, shipping costs

Why predict Urals loadings?
- Risk management: Forecast supply disruptions
- Trading: Arbitrage Brent-Urals differential
- Policy: Model sanctions effectiveness
- Hedging: Energy derivatives pricing

2020 challenge:
- Pre-pandemic: 13.6 mb/d (normal pre-pandemic level)
- Post-outbreak: oil market crash 3.5 mb/d low $\rightarrow$ 75% drop
- Duration: 2+ years (not temporary shock)
- Unprecendented in training data (2017-2019)

Backup: Feature definitions

Feature	Definition & rationale
Brent Futures	WTI/Brent price (USD/bbl) - key driver of margin
Urals Refining Crack	Refining margin (product price - crude price)
	Indicates refiner incentive to process Urals
MOEX Index	Russian stock market (reflects investor confidence)

All normalized using Z-score (mean 0, std 1) to account for scale differences

Backup: Model hyperparameters

Model	Architecture	Key hyperparameters
MLP	3-layer FC	Hidden units 32-16-8, dropout 25%, epoch 10
LSTM	Sequence model	Lookback 10, dropout 25%, epoch 10, batch 8
GBoost	Sequential trees	n_estimators 100, lr 0.1, depth 5
RF	Parallel trees	n_trees 100, max_depth 20, min_samples 5
SVM	RBF kernel	C=1.0, gamma=auto, kernel=rbf

Key finding

Even with regularization (dropout 25%), LSTM couldn't handle pandemic shock as well as MLP

Problem wasn't overfitting → it was architectural mismatch for regime-switching commodity data