Multi-Horizon Retail Demand Forecasting:

A Department-Level Pilot and a Four-Qubit Quantum Test

Rolling-origin evidence from log-transformed Salumeria department-level sales in an ordinary trading-day regime

EQAI 2026 - 5th European Summer School on Quantum AI

31 August - 04 September 2026

Lignano Sabbiadoro, Italy

Authors: ALESSANDRO MURATORE[1,2], GIUSEPPE AIELLO[1], SALVATORE QUARANTA[1], FILIPPO CAROLLO[1]

Affiliations: 1. Università degli Studi di Palermo, 2. PaghiPoco - Rocchetta S.R.L.

Email: alessandro.muratore01@unipa.it, giuseppe.aiello03@unipa.it, salvatore.quaranta@unipa.it, filippo.carollo@unipa.it

! KEY FINDING

RF leads at h=1; at h=7-28, leading classical losses are small relative to window variability. The tested 4-qubit VQC is not competitive in this pilot; matched PCA4 controls indicate unresolved compressed-learning effects.

1 ABSTRACT AND DATA

Abstract

Retail demand forecasting is tested on a real Salumeria department-level aggregated series using 604 observed trading days. The study separates an application benchmark from diagnostic controls: full-feature classical models are compared with a feasible four-qubit VQC pipeline, while matched-input controls help assess where performance deteriorates without claiming quantum advantage.

Observed data

- 604 observed trading days; 4 store-closure/major-holiday dates excluded.
- Single aggregated Salumeria department-level series; no point-of-sale identifier disclosed.
- Target: daily sales value, modelled as log(y) with y > 0.
- Three representative SKU profiles: one lumpy, two smooth.

SKU	Zero days	ADI	CV2	Class
Bauli	**51.5%**	**2.06**	**1.25**	**Lumpy**
Ham	5.8%	1.06	0.45	Smooth
Water	0.8%	1.01	0.42	Smooth

ADI = average demand interval; CV2 = squared coefficient of variation.

2 BENCHMARK PROTOCOL

Evaluation design

- Expanding-window monthly holdouts: Jul 2024-Aug 2025 (14 windows, N = 424 target days).
- Direct horizons: h = 1, 3, 7 and 28 trading days; closure/major-holiday dates excluded.
- Domain of evaluation: observed ordinary trading-day regime.
- Hyperparameters fixed ex ante and reused across all holdouts.
- MASE denominator: 7-trading-day seasonal naive scale; closures excluded.

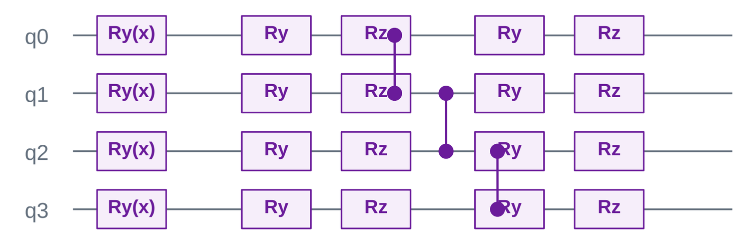
Features used

- 24 origin features: lagged sales, rolling means/SD, lagged promotion and price, origin calendar and trend.
- 5 known-ahead target calendar features only.
- Scaling and PCA are refitted on each training fold only.
- Lag and rolling features use only information available before the forecast target.
- No contemporaneous target-day sales or future promotion/price calendars are used.

Log(y) is used for scale compression; original-scale metrics use inverse transformation with smearing. RF ablation shows a modest, non-significant gain vs original scale (0.706 vs 0.737 MASE; p = 0.257).

3 4-QUBIT QUANTUM TEST

Implemented VQC architecture



Exact statevector simulator | 3 random seeds per window

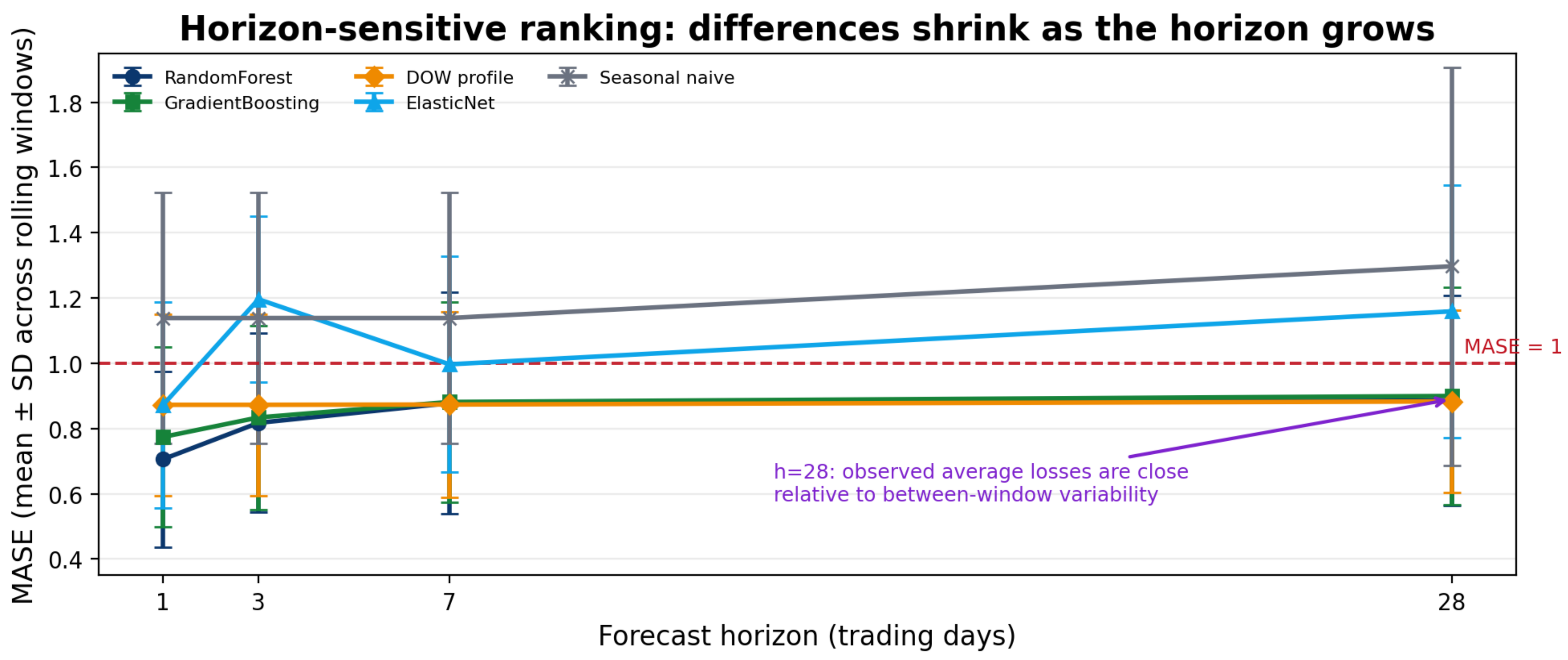
Input
29 engineered features -> PCA4 -> angle encoding

Circuit
4 qubits, 2 entangling layers, 16 circuit params

Readout
linear readout, 5 params; total 21 trainable params

The goal is not to claim quantum advantage or speed-up, but to test whether a small quantum embedding preserves useful forecasting information under four-qubit constraints.

4 MAIN RESULT: HORIZON-SENSITIVE RANKING



Finding: model complexity helps most at h=1. By h=7-28, observed differences among DOW, RF and GB are small relative to variability across rolling-origin windows; the forecast horizon matters more than the model label.

5 NUMERICAL SUMMARY

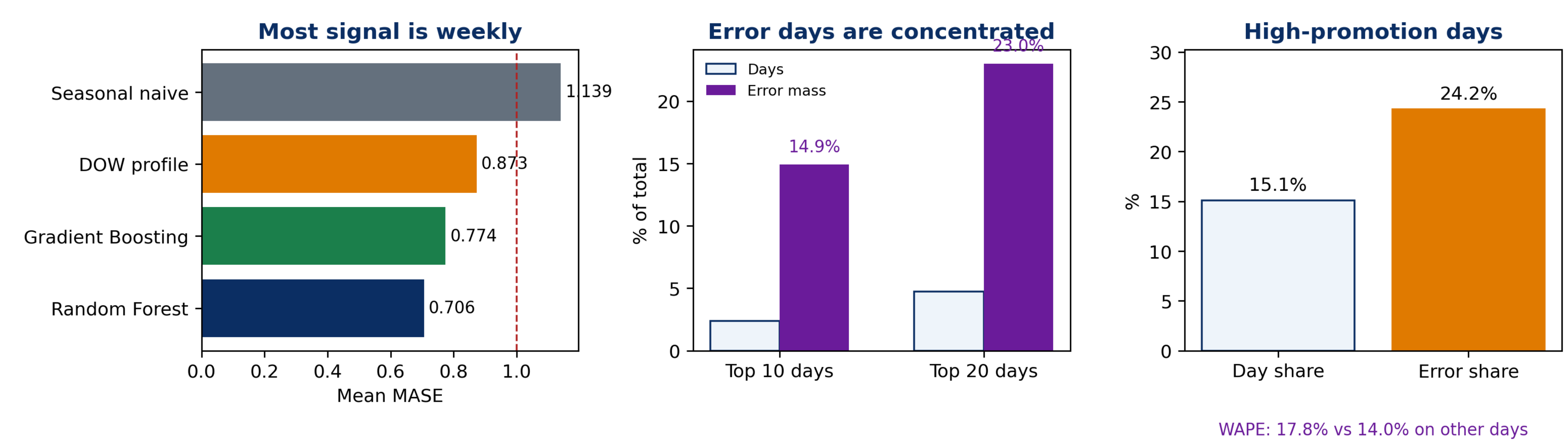
Horizon	Best/tie	MASE ± SD	WAPE	Comparator
h=1	**RF**	**0.706 ± 0.269**	**14.54%**	**DOW = 0.873**
h=3	RF	0.817 ± 0.275	16.98%	DOW = 0.873
h=7	**Tie**	**0.874-0.881**	**18.19-18.32%**	**DOW/RF/GB**
h=28	**Tie**	**0.883-0.900**	**18.37-18.65%**	**DOW/RF/GB**

Interpretation

- At short horizons, recent sales dynamics add value.
- At h >= 7, observed differences among leading classical models are small relative to between-window variability.
- Elastic Net was tested but did not win at h=28.
- This is why h=1 accuracy alone is insufficient for retail decisions.

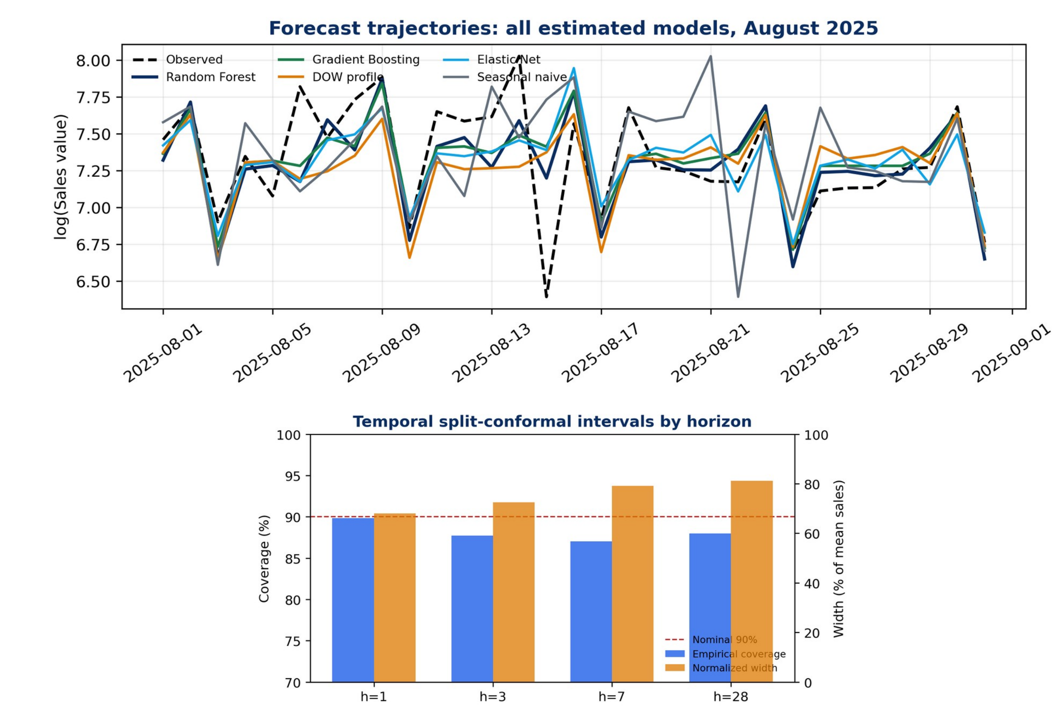
Practical rule: report uncertainty and horizon-specific performance, not only an overall winner.

6 WHAT DRIVES PREDICTABILITY AND ERROR?



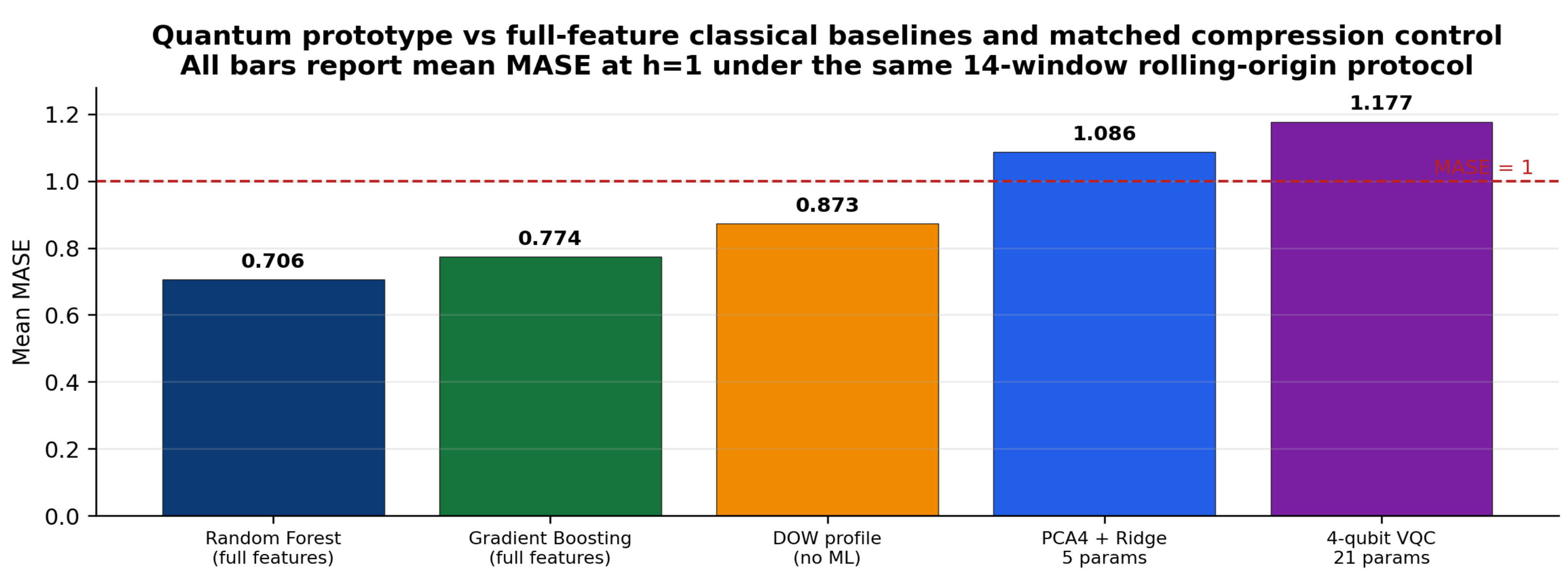
Diagnostic insight: the weekday profile is a strong no-ML baseline; the largest 20 error days produce 23.0% of total error; high-promotion days carry 24.2% of error on 15.1% of days, mainly due to higher volume rather than promotional unpredictability.

7 FORECAST TRAJECTORIES AND UNCERTAINTY



Temporal split-conformal intervals are approximate under serial dependence; coverage and width are reported by horizon and should be interpreted as empirical calibration diagnostics.

8 FOUR-QUBIT VQC RESULT: NEGATIVE IN THIS PILOT



PCA4 is unsupervised: 52% feature variance retained may preserve much less predictive signal. VQC adds parameters but degrades accuracy.

Result: the tested 4-qubit VQC does not outperform either full-feature classical baseline (RF 0.706; GB 0.774). PCA4 + Ridge indicates that performance deteriorates before or during the compressed-learning stage, but additional matched-input controls are needed to distinguish information loss, encoding, circuit capacity and optimization effects.

9 DIAGNOSIS AND VALIDITY THREATS

Ridge indicates deterioration before or during the compressed-learning stage, but additional matched-input controls are needed to distinguish information loss, encoding, circuit capacity and optimization effects.

Diagnostic interpretation

- PCA4 is unsupervised and retains only 52.0% of feature variance; retained variance may not correspond to retained predictive signal.
- PCA4 + Ridge uses 5 parameters and already performs poorly (MASE 1.086).
- The VQC uses 21 parameters yet degrades to MASE 1.177, consistent with possible optimization difficulty or poorly conditioned regions of the loss landscape.
- RF (0.706) and GB (0.774) define the full-feature classical performance envelope.
- No quantum speedup, advantage or general QML failure is claimed.

Validity limits

- One department-level pilot series: external validity is limited, despite 14 temporal windows.
- Only representative SKUs are available for ADI/CV2 diagnosis.
- Major holiday closure dates are excluded; results describe the ordinary trading-day regime.
- Known-ahead promotion calendars and inventory outcomes are not yet included.
- Statevector results must be replicated with shot noise and hardware constraints.

Strength: evidence, classical baselines and quantum hypotheses are explicitly separated.

10 CONCLUSIONS

- Forecasting performance is horizon-dependent: h=1 accuracy does not guarantee replenishment relevance.
- Weekly structure is a strong transparent predictor; ML adds most value at short horizons.
- The tested four-qubit VQC is not competitive in this pilot, but the negative result is experimentally useful.
- Temporal conformal intervals are approximate; operational decisions require horizon-specific quantiles.

Central message: a poster-worthy Quantum AI result can be negative when benchmark performance is separated from unresolved representation and optimization effects.

11 NEXT EXPERIMENTS

- Confirm the real replenishment lead time and prioritize that horizon.
- Evaluate h-specific quantile forecasts with pinball loss and interval score.
- Add matched-input nonlinear controls (RF/GB/MLP on PCA4) and supervised compression.
- Replicate VQC with more seeds, gradient diagnostics, shot noise and hardware-constrained runs.
- Run inventory simulation before claiming operational impact.

12 REPRODUCIBILITY CHECKLIST

- 604 trading days documented; 4 closure/major-holiday dates excluded in every table.
- 29 feature protocol and lag structure specified.
- 14 rolling-origin windows fixed; scaling and PCA refitted inside each training fold.
- Lag/rolling features use pre-target information only.
- MASE/WAPE reported on original scale after inverse log transformation; MAE/RMSE on log scale when shown.
- Raw data confidential; derived code and metrics can be shared.

Business impact remains potential until linked to lead time and inventory simulation.

REFERENCES

- [1] Biamonte, J. et al. Quantum machine learning. Nature 549, 195-202 (2017).
- [2] Syntetos, A.A., Boylan, J.E. & Croston, J.D. On the categorization of demand patterns. JORS 56, 495-503 (2005).
- [3] Friedman, J.H. Greedy function approximation: a gradient boosting machine. Ann. Stat. 29, 1189-1232 (2001).
- [4] Hyndman, R.J. & Koehler, A.B. Another look at measures of forecast accuracy. IJF 22, 679-688 (2006).
- [5] Duan, N. Smearing estimate: a nonparametric retransformation method. JASA 78, 605-610 (1983).
- [6] Havlicek, V. et al. Supervised learning with quantum-enhanced feature spaces. Nature 567, 209-212 (2019).
- [7] Schuld, M. et al. Evaluating analytic gradients on quantum hardware. Phys. Rev. A 99, 032331 (2019).
- [8] Angelopoulos, A.N. & Bates, S. Conformal prediction: a gentle introduction. FnT ML 16, 494-591 (2023).
- [9] Barber, R.F., Candes, E.J., Ramdas, A. & Tibshirani, R.J. Conformal prediction beyond exchangeability. Ann. Stat. 51, 816-845 (2023).

Bibliographic metadata checked against publisher or primary-source records.

We thank the retail partner for data access and operational context.

Contact:
Corresponding contact: alessandro.muratore01@unipa.it