Lecture 12: Capstone & Applications

Synthesising the Forecasting Toolkit

BSAD 8310: Business Forecasting

University of Nebraska at Omaha

Spring 2026

1 Course Synthesis

2 Combining and Testing Forecasts

3 RSXFS Final Leaderboard

4 Case Study: Utility Demand Forecasting

5 Communication and Deployment

6 Takeaways and References

Course Synthesis

Twelve lectures, two frameworks, one decision: which tool fits this problem?

Part I: Classical Forecasting (L01–L06)

- **L01** Benchmarks and evaluation discipline
- **L02** Regression-based forecasting
- **L03** Exponential smoothing (ETS)
- **L04** ARIMA and Box-Jenkins workflow
- **L05** Multivariate: VAR, ARIMAX, cointegration
- **L06** Forecast evaluation, DM test, combination

Part II: Machine Learning (L07–L11)

- **L07** Bias-variance, train/val/test, CV
- **L08** Regularization: LASSO, Ridge, Elastic Net
- **L09** Tree methods: Random Forests, XGBoost
- **L10** Neural networks: LSTM, attention
- **L11** Feature engineering and pipeline design

Capstone question: Given a new forecasting problem, which part of the toolkit do you reach for first — and how do you decide when to switch?

1. **How much data?** $n < 200$: classical methods are more reliable; $n \geq 200$: either framework is viable.
2. **How many predictors?** $k < 10$: ARIMAX or VAR; $k \geq 10$: regularization or trees.
3. **Is the pattern regular?** Strong, stable seasonality $\Rightarrow$ SARIMA/ETS are competitive.
4. **Interpretability required?** Yes $\Rightarrow$ LASSO or SARIMA; No $\Rightarrow$ XGBoost or LSTM.
5. **Refit cadence?** Weekly $\Rightarrow$ prefer simpler models; Monthly $\Rightarrow$ ML feasible.

No single model wins every case. The framework replaces intuition with discipline. Apply it before fitting.

1. **Data leakage** — features using future information inflate in-sample accuracy and collapse out-of-sample. (Fix: `.shift(1)` before every rolling window.)
2. **In-sample evaluation only** — training RMSE is not forecast accuracy. Always evaluate on a held-out test set.
3. **Wrong metric for the cost structure** — MAPE fails near zero; RMSE penalizes outliers heavily. Match metric to business consequences. (L01, L06)
4. **No statistical test for differences** — a lower RMSE may be noise. Report Diebold–Mariano p-values. (L06)
5. **Point forecast without uncertainty** — a forecast without a prediction interval is not actionable for most business decisions. Report intervals.

Combining and Testing Forecasts

Equal-weight combination is the benchmark that beats most individual models.

(Bates and Granger 1969; Timmermann 2006; Stock and Watson 2004)

Why combination works: individual models capture different features of the DGP. Combining reduces variance without increasing bias.

RSXFS result:

LSTM alone: RMSE $= 1{,}920$
XGBoost alone: RMSE $= 2{,}050$
SARIMA alone: RMSE $= 2{,}840$

Equal-weight (SARIMA + XGBoost + LSTM): RMSE $= 2{,}080$ — within 8% of LSTM, lower variance, simpler to maintain.

The top-ranked hybrid (ES-RNN) used combination of exponential smoothing with an RNN internally.

Pure ML without combination ranked *lower* than Theta (classical) on short series. (Makridakis et al. 2020)

(Diebold and Mariano 1995; Harvey et al. 1997)

Pairwise Diebold–Mariano test results on RSXFS (walk-forward errors, HAC standard errors). $\star p < 0.05$; $\star\star\, p < 0.01$; $\star\star\star p < 0.001$; n.s. not significant.

	vs. SARIMA	**vs. Elastic Net**	**vs. RF**	**vs. XGBoost**
Elastic Net	$\star\star\star$	—		
Random Forest	$\star\star\star$	$\star\star$	—	
XGBoost	$\star\star\star$	$\star\star\star$	$\star$	—
LSTM	$\star\star\star$	$\star\star\star$	$\star\star$	n.s.
Combination	$\star\star\star$	$\star\star\star$	$\star$	n.s.

LSTM vs. XGBoost: **not significant** (n.s.). The RMSE gap ($1{,}920$ vs. $2{,}050$) does not clear the DM threshold. Report p-values, not just RMSE gaps.

RSXFS Final Leaderboard

Eleven methods, one dataset, a clear pattern.

Lecture	Model	RMSE	MAE
L01	Seasonal Naïve (benchmark)	4,210	3,120
L03	ETS (auto-AIC)	2,890	2,150
L03	Holt-Winters (add.)	2,950	2,190
L04	SARIMA$(1,1,1)(1,1,1)_{12}$	2,840	2,100
L05	ARIMAX ($+$ sentiment index)	2,780	2,060
L08	Elastic Net†	2,410	1,800
L08	Ridge†	2,460	1,830
L09	Random Forest†	2,210	1,640
L09	XGBoost†	2,050	1,510
L10	LSTM (2-layer, $T = 24$)†	1,920	1,410
L06	Equal-weight combination	2,080	1,530

Equal-weight combination (2,080) beats XGBoost (2,050) on MAE and nearly matches LSTM — at a fraction of the deployment complexity.

Statistically confirmed improvements:

- All ML methods $\gg$ SARIMA ($p < 0.001$)
- XGBoost $>$ Elastic Net ($p < 0.001$)
- RF $>$ Elastic Net ($p < 0.01$)
- XGBoost $>$ RF ($p < 0.05$) — marginal

Not statistically confirmed:

- LSTM vs. XGBoost (n.s.)
- Combination vs. XGBoost (n.s.)
- Ridge vs. Elastic Net (n.s.)

Practical decision rule:

If two models are DM-indistinguishable, choose the *simpler* one.

LSTM $\approx$ XGBoost (DM n.s.) $\Rightarrow$ prefer XG-Boost: fewer hyperparameters, faster refit, more interpretable feature importance.

Five-step production pipeline:

Fit SARIMA as monitoring anchor and classical baseline. Refit monthly on expanding window.

Fit XGBoost with 36-feature set. Retrain monthly; monitor feature drift.

Form equal-weight combination: $\hat{y}_t = \frac{1}{2}(\hat{y}_t^{\text{SARIMA}} + \hat{y}_t^{\text{XGB}})$.

Monitor quarterly: run DM test on trailing 12 months. If ML no longer improves on SARIMA ($p > 0.10$), revert to SARIMA until data accumulates.

Report intervals: 80% and 95% prediction intervals from bootstrap over walk-forward residuals.

Combination RMSE (2,080) vs. XGBoost alone (2,050): the 30-unit RMSE gain from LSTM does not justify the added deployment complexity for most business settings.

Case Study: Utility Demand Forecasting

A different domain to stress-test the decision framework.

Dataset: Monthly U.S. residential natural gas consumption (RESGAS, EIA/FRED series `NGRESCON`).

- **Period:** Jan 2005 – Dec 2023 ($n = 228$)
- **Train:** 2005–2019; **Test:** 2020–2023
- **Seasonality:** extreme (6× winter/summer ratio), virtually no AR structure beyond $m = 12$
- **Trend:** slow decline (efficiency gains)

Business stake: Natural gas procurement, pipeline capacity planning, and hedging contracts require 12-month-ahead forecasts. Errors translate directly to over-purchase costs or supply shortfalls.

Test period challenge: COVID-19 (2020) shifted residential usage patterns. The test period is adversarial — an honest stress test for all methods.

Contrast with RSXFS: retail sales are driven by consumer demand shocks; natural gas demand is driven by weather and long-term efficiency trends.

Q	Question	Answer for RESGAS	Implication
1	Series length?	228 months	Either framework
2	Predictor count?	≤ 5 core features	Classical competitive
3	Seasonality regular?	**Yes** — almost purely sinusoidal	SARIMA/ETS strong
4	Interpretability?	Regulatory context	Prefer classical
5	Refit cadence?	Monthly feasible	ML possible

Framework prediction (before fitting): SARIMA and ETS will be competitive with ML. Feature engineering will add less value than on RSXFS. Prefer SARIMA for interpretability and audit compliance.

Test-set RMSE (2020–2023, walk-forward, units: billion cubic feet/month):

Model	RMSE
Seasonal Naïve	12,400
ETS	4,650
SARIMA	**4,200**
Elastic Net	5,800
Random Forest	5,100
XGBoost	4,900
LSTM	4,750

SARIMA wins (RMSE = 4,200) — the decision framework correctly predicted the outcome before a single model was fitted.

ML methods (LSTM best at 4,750) do not overcome the regular seasonality that SARIMA captures exactly.

Contrast with RSXFS: LSTM = 1,920 vs. SARIMA = 2,840. The feature gap from L11 exists only when the series has exploitable nonlinear structure.

Transferred from RSXFS workflow:

- Walk-forward CV discipline (no shuffling)
- Out-of-sample test set (held out throughout)
- DM test for statistical significance
- Forecast combination as a fallback
- Prediction interval reporting

Did not transfer:

- ML's large RMSE advantage over SARIMA
  (regular seasonality negates the feature advantage)
- 36-feature set value
  (gas demand has 5 core drivers, not 36)
- LSTM's edge over XGBoost
  (sequential patterns already captured by lags)

The decision framework worked. Applying five questions before fitting directed resources toward SARIMA, which won. The framework is the product — not any specific model.

Communication and Deployment

A forecast without uncertainty is just a guess in formal notation.

What to include:

- Point forecast with 80% and 95% prediction intervals
- Model used, training window, last refit date
- RMSE on most recent test period
- Known limitations and assumption violations
- Update cadence and trigger for review

What to exclude:

- In-sample R^2 or training RMSE
- p-values without business interpretation
- Decimal precision beyond measurement error

RSXFS 12-month forecast
Point estimate: $452,000M
80% PI: $444,000M – $460,000M
95% PI: $438,000M – $466,000M

Model: XGBoost + SARIMA combination, retrained monthly (36 features).
Test RMSE (2020–2023): $2,080M.

Next update: April 2026 after March data release.

Four monitoring checks:

1. **Rolling RMSE:** compute walk-forward RMSE each period. Flag if $> 2\times$ historical average over three consecutive months.
2. **Quarterly DM test:** confirm ML still outperforms SARIMA baseline. Revert if $p > 0.10$.
3. **Feature drift:** monitor input feature distributions. Alert if mean shifts $> 2\sigma$ from training distribution.
4. **Model refresh:** retrain on expanding window monthly. Benchmark against prior version before deploying.

Models degrade silently.

A model that was accurate in 2022 may be systematically wrong in 2025 due to structural shifts in the economy.

A monitoring dashboard is not optional — it is the difference between a deployed model and a science project.

Takeaways and References

What we learned and where to go next.

Classical methods (ETS, ARIMA) are competitive for regular, short series with few predictors. Match the method to the series structure.

ML methods (regularization, trees, LSTM) add value when features are rich, series are long, and patterns are nonlinear or structural.

Forecast combination (equal-weight) consistently matches or beats the best individual model at lower deployment variance (Bates and Granger 1969; Timmermann 2006).

Evaluation discipline — walk-forward CV, DM significance test, out-of-sample only — is non-negotiable (Diebold and Mariano 1995).

Feature engineering often yields larger RMSE gains than switching model class. The feature gap dominates the model gap (Makridakis et al. 2020).

Communication determines whether a technically correct forecast is actionable. Report intervals, not just point estimates.

This concludes BSAD 8310: Business Forecasting. The toolkit, the discipline, and the decision framework are yours.

- Bates, John M. and C. W. J. Granger (1969). "The Combination of Forecasts". In: *Operational Research Quarterly* 20.4, pp. 451–468.
- Diebold, Francis X. and Roberto S. Mariano (1995). "Comparing Predictive Accuracy". In: *Journal of Business & Economic Statistics* 13.3, pp. 253–263.
- Harvey, David, Stephen Leybourne, and Paul Newbold (1997). "Testing the Equality of Prediction Mean Squared Errors". In: *International Journal of Forecasting* 13.2, pp. 281–291.
- Makridakis, Spyros, Evangelos Spiliotis, and Vassilios Assimakopoulos (2020). "The M4 Competition: 100,000 Time Series and 61 Forecasting Methods". In: *International Journal of Forecasting* 36.1, pp. 54–74.
- Stock, James H. and Mark W. Watson (2004). "Combination Forecasts of Output Growth in a Seven-Country Data Set". In: *Journal of Forecasting* 23.6, pp. 405–430.
- Timmermann, Allan (2006). "Forecast Combinations". In: *Handbook of Economic Forecasting* 1, pp. 135–196.