Sales Prediction for Wellington Management

Group members: Zhicheng He, Louren Park, Daniel McClafferty

Industry executive: Yossi Cohen, Wellington Management

Advisor: Professor Ali Hirsa, Columbia University IEOR

Background and Introduction

Forecasting firm-level revenue growth is central in applied finance but highly noisy due to macroeconomic and firm-specific drivers.

Past Wellington projects showed advanced deep learning (e.g., CNN–Attention LSTMs) can capture nonlinear dynamics but struggle to generalize.

Our project builds on this foundation by:

- Expanding to 578 Bloomberg macro time series (2010–2025).
- Refining preprocessing pipelines for stability.
- Comparing regression, recurrent nets (GRU/LSTM), and tree ensembles (XGBoost).

Goal: identify models that best balance accuracy, robustness, and interpretability.

Data Collection

Firm Sales

- 51 public firms, monthly sales.
- Target: YoY growth; added lags; forward-filled gaps.

Macroeconomic Indicators

- 578 Bloomberg series across labor, housing, trade, production, sentiment.
- Cleaned + standardized; reduced to ~120 robust features.
- Used both actual values and "surprise" (Actual – Survey).

Integrated Dataset

- Panel: firm × month with sales + macro drivers.
- Provides broad foundation for testing models.

Data Preprocessing

To improve model performance and interpretability, we applied three tailored preprocessing strategies:

VIF for Traditional Models

For traditional models (Linear, Lasso, Ridge), we used Variance Inflation Factor (VIF) filtering to mitigate multicollinearity:

- Retained features with VIF < 5
- Ensures interpretability and stable coefficients without redundant predictors

PCA for Advanced Models

We used Principal Component Analysis (PCA) to reduce dimensionality for deep learning models. The top 25 components captured over 95% of total variance:

- PC1: Employment indicators (e.g., nonfarm payrolls)
- PC2: Inflation expectations and labor sentiment
- PC3–PC5: Housing, GDP, and manufacturing trends

PCA helps LSTM models focus on major macroeconomic signals and reduce noise

Feature Selection via Spearman Correlation

To prioritize macro indicators, we computed quarterly Spearman rank correlations with firm-level YoY sales growth.

- Captures monotonic (not just linear) relationships.
- Averaged across quarters to smooth seasonality.
- Ranked by absolute correlation to highlight strongest signals.

Top indicators:

- Positive → trade balance, Richmond Fed manufacturing survey.
- Negative → CPI momentum, existing home sales.

These features informed input design for GRU and XGBoost models.

Traditional Models

OLS Regression

- Train R^2 = 0.73, MAE ≈ 1274
- spend_amount_agg most influential, embedding features (embed_1, embed_2, embed_5) also strong
- Underestimates extreme sales, residuals show heteroskedasticity

Lasso Regression

- Performance similar to OLS (R^2 = 0.73, MAE ≈ 1273)
- Enforces sparsity: weak lagged ratios shrunk toward zero
- Improves interpretability, but predictive accuracy unchanged

Ridge Regression

- Performance comparable (R^2 = 0.73, MAE ≈ 1274)
- Stabilizes coefficients in presence of multicollinearity
- Retains more features than Lasso but still weak for extreme outcomes

Takeaway

- All three regressions plateaued at R^2 ≈ 0.73
- Regularization (Lasso, Ridge) improved coefficient stability/interpretability but did not enhance predictive accuracy
- Linear models fail to capture nonlinear and volatile sales dynamics → motivates advanced methods

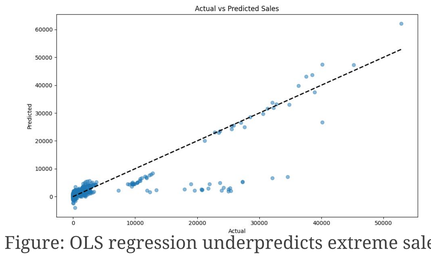


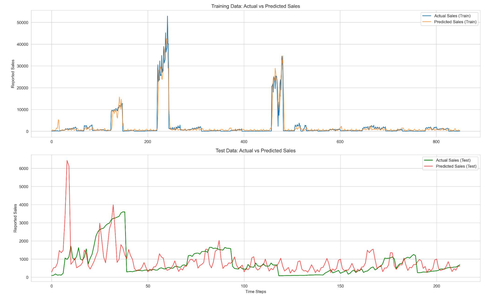
Figure: OLS regression underpredicts extreme sales

Advanced Models

Baseline LSTM Model

- Trained on PCA-transformed macro features using a sliding 3-month window to predict next-month sales.
- Architecture: 2 LSTM layers (64, 32 units) → Dense (16 ReLU) → Output.
- Performance:
  - Train R^2 = 0.94, strong in-sample fit
  - Test R^2 = –0.12, weak generalization, especially around sales spikes

Highlights the need for better generalization strategies.



Percentile-based LSTM Models

- Split data into 5 sales quantiles (Q1–Q5), trained models to predict relative sales level.
- Best result: Q1 test R^2 = 0.59, Worst result: Q5 test R^2 = –5.44

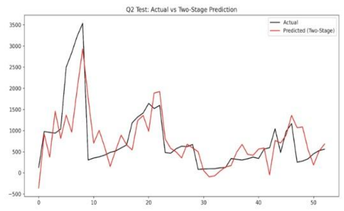
Effective for low sales, but unstable in high-volume segments due to skewed distribution.

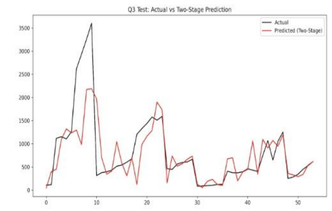
Quarter-based LSTM Models

- Separate models for Q1–Q4 to capture seasonality
- Train R^2: 0.64–0.80 | Test R^2: 0.59–0.79
- Strongest test performance in Q3 (R^2 = 0.79)
- Improves generalization vs. baseline (Test R^2 = –0.12)

Two-Stage Quarterly LSTM + XGBoost

- Train R^2 ≈ 1.00 | Test R^2 varies by quarter
- Best generalization in Q2 (R^2 = 0.33) and Q3 (R^2 = 0.54)
- Q1 & Q4 overfitted: poor test R^2 (–6.31, –1.19)
- Highlights sensitivity to seasonal volatility



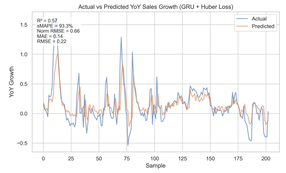


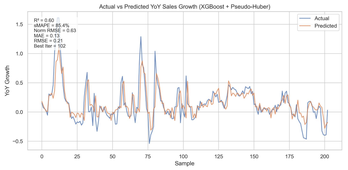
GRU Model with Spearman Features

- Inputs: lagged sales growth + top macro indicators (ranked by Spearman correlation).
- Architecture: 1 GRU layer → Dense output, trained with Adam + pseudo-Huber loss (δ=1) for outlier robustness.
- Performance:
  - R^2 = 0.57, sMAPE = 93%, Normalized RMSE = 0.66, MAE = 0.14, RMSE = 0.22
- Takeaway: Adding macro signals improves accuracy over an autoregressive baseline, but errors remain large during extreme fluctuations.

Tree-Based Model (XGBoost)

- Inputs: lagged sales growth + Spearman-selected macro deltas.
- Config: Pseudo-Huber objective, depth=3, η=0.05, early stopping (102/300 rounds).
- Performance:
  - R^2 = 0.60, sMAPE = 85%, Normalized RMSE = 0.63, MAE = 0.13, RMSE = 0.21
- Takeaway: Slightly outperforms GRU, handling nonlinearities and cross-sectional noise better. Suggests value in GRU+XGBoost ensembles.





Conclusion

We built a robust macro–micro forecasting pipeline that integrates macroeconomic signals with firm-level sales data. Quarter-specific LSTM models improved generalization by capturing seasonal patterns, while two-stage models enhanced training accuracy but showed limited gains out-of-sample. Simpler models like XGBoost and GRU offered strong, stable baselines. Future work can further refine these approaches to improve robustness and forecasting power.

Future Work

- **Expand the dataset:** Include more firms and longer historical periods, or use synthetic data to capture diverse market conditions and stress-test model robustness.
- **Broaden evaluation metrics:** Go beyond R^2 and MAE by incorporating RMSE, MAPE, AUC, and other metrics to better assess model performance.
- **Add a validation set:** Introduce a dedicated validation split to improve hyperparameter tuning and ensure more reliable generalization.