Week 6 Report

Sales Forecasting Based on Past Sales Statistics

Nguyen Minh Duc - 20225437

Week 6 Objectives:

- Attempt implementing Multivariate GRU and compare against Univariate GRU.
- Compare and perform Error Analysis for implemented models.

---

1 Multivariate GRU Experiment

1.1 Motivation and summary

This section documents a controlled attempt to extend the previously-developed univariate GRU (which used only past sales values) into a multivariate model that ingests engineered autoregressive and rolling-statistic features. The goal is to test whether additional, causally valid features improve forecasts under the same settings.

1.2 Implementation details

While the overall experimental protocol was kept identical to the univariate GRU for comparability, several implementation aspects necessarily diverged when extending to the multivariate setting.

- **Input dimensionality.** Instead of a univariate input sequence of shape $(seq\_len, 1)$, the multivariate GRU consumes sequences of shape $(seq\_len, n\_features)$, where $n\_features$ includes the target and engineered features (e.g., lagged values and rolling statistics).
- **Feature construction and availability.** Engineered features contain missing values at the beginning of the time series (e.g., lag-4 or rolling-window statistics). This issue is absent in the univariate case and requires explicit handling to prevent NaNs from propagating into training sequences.
- **Preprocessing complexity.** Scaling is performed feature-wise using a multivariate StandardScaler fitted on training sequences at each walk-forward step. In contrast, the univariate GRU scales a single channel and is unaffected by cross-feature NaN contamination.
- **Failure modes and safeguards.** The multivariate pipeline is more susceptible to degenerate training states (e.g., zero valid sequences after NaN filtering). Additional safety checks and deterministic fallbacks were implemented.

1.3 Experiments and results

We evaluated the model using a small set of feature lists. The features used include lagged and roll-mean versions of total sales, which are directly correlated to the goal feature. The table below summarizes the metrics computed using different subsets of the feature list.

Table 1: Phase 1 Multivariate GRU results (selected feature sets).

Feature set	**RMSE**	**MAPE (%)**	**ME**
`["total_sales"]` (univariate GRU baseline)	42 155.48	9.6451	-4 069.99
`["total_sales", "total_sales_lag_1"]`	43 254.50	9.9376	-3 893.53
`["total_sales", "total_sales_lag_1", "total_sales_lag_4"]`	51 224.63	12.1684	5 799.10
`["total_sales", "total_sales_lag_4"]`	48 178.30	11.2332	4 985.82
`["total_sales", "total_sales_lag_1", "total_sales_rollmean_4"]`	51 296.74	11.3967	1 142.97
`["total_sales", "total_sales_lag_4", "total_sales_rollmean_4"]`	47 220.55	10.8655	4 912.26

Observations

- The **univariate GRU** (only `total_sales`) produced the best and most stable results among the GRU variants tested.
- Adding seemingly-relevant engineered features (lags and short rolling statistics) degraded performance in multiple configurations. In many cases, both RMSE and MAPE increased, and ME increased in magnitude, indicating increased bias or instability rather than improvement.
- The degradations are consistent with the hypothesis that the GRU already learns effective short-term dynamics from raw sequences; explicitly adding the same signals as separate features increases model input redundancy and raises training variance under the small-sample, walk-forward regime.

1.4 Interpretation and explanation

The empirical behaviour can be understood from both modelling and optimization perspectives:

1. **Redundancy.** Lags and rolling statistics are deterministic transforms of the target; a recurrent model with sufficient capacity implicitly forms lagged representations in its hidden state. Adding those engineered transforms therefore provides little new information but increases input dimensionality and collinearity.

2. **Small-sample, high-variance training.** Walk-forward retraining with a compact dataset produces few sequences for early folds; additional inputs increase gradient variance and make optimization less stable. This explains the large swings in ME and occasional catastrophic increases in RMSE.

3. **Numerical conditioning and NaN propagation.** Because rolling features are undefined at the start of the series, sequences can contain NaNs that, if not handled, corrupt scaler statistics and model parameters. Although this was mitigated by causal imputation and sequence-dropping, the presence of initially-missing values amplifies the fragility of multivariate training.

Although the Multivariate GRU could not improve results, this experiment provided important insights in model design and evaluation, including null value handling, the information gained by engineered features for recurrent networks, etc.

2 Model Selection

This section consolidates the performance of all forecasting models implemented so far under a unified walk-forward evaluation protocol. To maintain clarity and focus, only models that demonstrated competitive or representative behavior are included. Multivariate GRU configurations are excluded, as none consistently outperformed the univariate GRU baseline.

2.1 Quantitative Comparison

Table 2 summarizes the predictive performance of all selected models using RMSE, MAPE, and Mean Error (ME). Lower RMSE and MAPE indicate better accuracy, while ME reflects systematic bias.

2.2 Model-wise Analysis

Naive and Seasonal Baselines. The naive and seasonal naive models serve as reference points. As expected, their performance is substantially weaker than more structured approaches, highlighting the presence of non-trivial temporal dynamics beyond simple persistence or fixed seasonality.

Table 2: Performance comparison of forecasting models

Model	RMSE	MAPE (%)	ME
Naive	63 428.41	13.90	-664.72
Moving Average (4 weeks)	43 698.10	9.99	-1 325.87
Seasonal Naive (4 weeks)	52 395.30	12.14	-2 456.74
ARIMA (1,1,1)	54 110.44	11.81	1 239.32
Prophet	52 868.51	13.96	28 891.20
LightGBM	46 722.61	10.57	-6 211.58
XGBoost	44 282.17	9.82	-1 864.16
Univariate GRU	**42 155.48**	**9.65**	-4 069.99

Moving Average. The moving average baseline performs surprisingly well, achieving competitive RMSE and MAPE relative to more complex models. This indicates that short-term temporal smoothing captures a significant portion of the signal in weekly sales data, and sets a strong baseline that more advanced models must exceed to justify added complexity.

ARIMA. ARIMA (1,1,1) does not outperform simpler baselines in this setting. While it captures local temporal dependencies, its limited flexibility and sensitivity to non-stationarity likely constrain its effectiveness on this dataset.

Prophet. Prophet underperforms relative to most other models, particularly in MAPE and ME. The large positive bias suggests systematic overestimation, possibly due to mismatches between Prophet's trend/seasonality assumptions and the observed sales dynamics.

LightGBM. LightGBM demonstrates strong predictive power among tree-based models, but exhibits a noticeable negative bias. This suggests that while it captures variance reasonably well, it may systematically underpredict peaks, potentially due to conservative splitting behavior or regularization effects.

XGBoost. XGBoost improves upon LightGBM in both RMSE and MAPE, with reduced bias. Its stronger performance is consistent with its more aggressive boosting strategy and robustness to feature interactions, making it the most effective non-neural model tested so far.

Univariate GRU. The univariate GRU achieves the best overall RMSE and lowest MAPE among all evaluated models. Despite a moderate negative bias, it shows improved alignment with temporal patterns and peak movements, indicating that sequence modeling provides tangible benefits even without exogenous features.

2.3 Interim Conclusions

From a weekly reporting perspective, several conclusions can be drawn:

- Strong baselines (moving average) remain difficult to beat, emphasizing the importance of careful benchmarking.
- Tree-based models offer meaningful gains over classical statistical models, with XGBoost outperforming LightGBM in this study.
- Sequence models (univariate GRU) provide the strongest performance so far, but gains over XGBoost are incremental rather than transformative.

These findings motivate further refinement of neural sequence models and targeted error analysis, which will be addressed in subsequent weeks.