Lightweight Multimodal Ensemble for GOLD-Price Forecasting

By: *Abhinav Vadhera, Rodrigo Lopez, & Vikyath Naradasi*

CSCI 567: Machine Learning

Prof: *Victor S. Adamchik*, TA: *Philipp Eibl*

TABLE OF CONTENTS

01

Problem Statement

Motivation

- **Gold** volatility surged post-2020, driven by macroeconomic events.
- Sensitive to CPI inflation, USD fluctuations, and oil market volatility.
- Traders and institutions need short-term forecasts for hedging.
- Ideal candidate for multivariate time-series ML prediction.
- **High market value:** even small forecasting improvements offer significant returns

Dataset

- Collected daily data from Yahoo Finance, Kaggle, & FRED (2010–2025).
- Series:
  - Gold (GC=F),
  - USD Index (DX-Y.NYB),
  - Oil (CL=F),
  - CPI.
- Engineered 11 features: returns, moving averages:
  - **(5-day, 20-day), volatility, RSI.**
- Min-max scaling applied to standardize features (0–1 range).
- Sliding-window: 30-day inputs → next-day gold prediction.
- Final split: 3045 training samples, 760 validation samples.

02

Models

A MULTIMODAL APPROACH

BI-GRU

1-layer, 32 units; captures sequential patterns.

TRANSFORMER

HuggingFace Tiny; attention identifies critical days.

TCN

Dilated causal convolutions (3 layers); efficient and fast.

- ❑ Each model trained individually on identical dataset.
- ❑ Model diversity ensures complementary strengths/weaknesses.
- ❑ Allows ensemble to capitalize on differing prediction patterns.

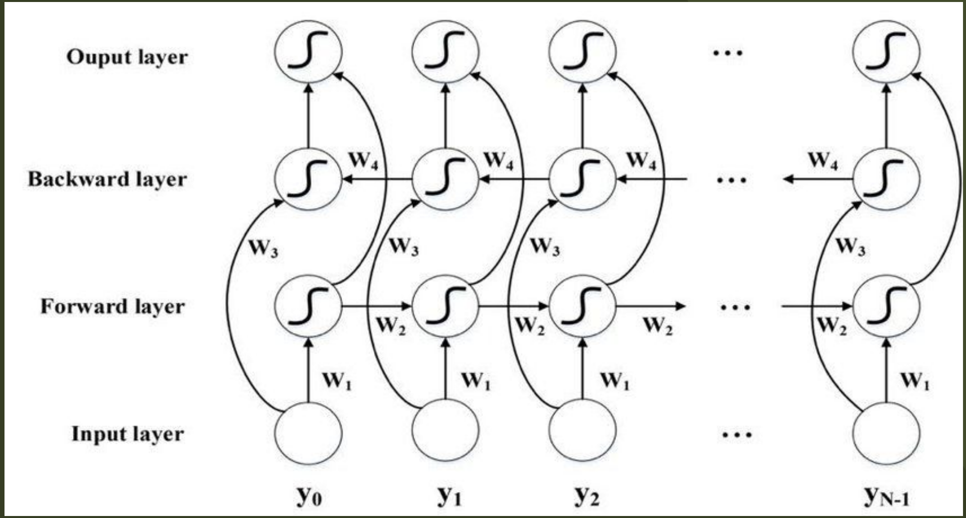
Bi-GRU (Bi-GRU) Overview

- Bidirectional Gated Recurrent Unit:
  - Specialized RNN that captures sequential relationships.
  - Uses gates to control information flow (update/reset gates).
- 2 GRUs stacked: one processes data forward (past→future), another backward (future→past).
  - Captures context from both past and future data points.
  - Ideal for time-series data where both previous and subsequent context matter.

- Why did we use Bi-GRU?
  - Excellent for short-term memory:
    - remembers key market events.
  - Lightweight:
    - few parameters, quick to train.
  - Naturally handles time-series:
    - Gold prices strongly dependent on historical context!!

Bi-GRU (Bi-GRU) Overview

- Our Bi-GRU Setup
  - Single-layer Bidirectional GRU.
  - 32 hidden units per direction (forward/backward).
  - Input:
    - 30-day window of market and indicator features.
  - Output:
    - Prediction of gold price (next-day value, scaled).



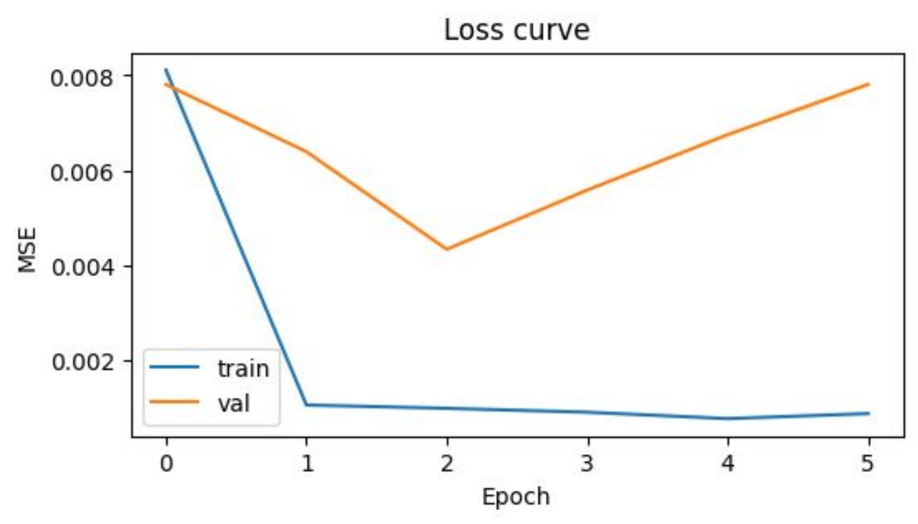
Bi-GRU Results

Provided robust baseline for ensemble fusion = **(71% weight in final ridge ensemble)**

Time taken to train on T4 GPU = **~290 secs**

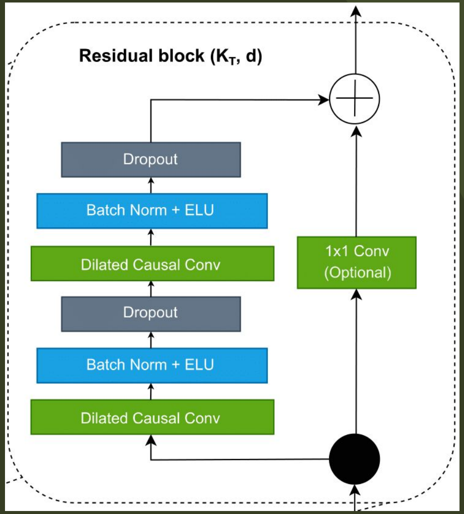
Best single-model RMSE = **0.067 (scaled)**

Bi-Direction accuracy = **~49%**



Temporal Convolutional Network (TCN) Overview

- TCN Concept:
  - - CNN adapted for time series prediction
  - - Processes multiple time steps in parallel (vs sequential RNNs)
  - - Uses "causal" convolutions - NO future data leakage
- Multi-Scale Pattern Detection:
  - - Dilated convolutions see different time scales simultaneously
  - - Dilation 1: Daily patterns [day t-2, t-1, t]
  - - Dilation 2: Weekly patterns [day t-4, t-2, t]
  - - Dilation 4: Monthly patterns [day t-8, t-4, t]
- Key Advantages:
  - - Faster training: Parallel processing
  - - Long memory: Large receptive field
  - - Interpretable: Can visualize learned filters
  - - Causal: Respects temporal order



TCN Code Implementation Details

- Architecture We Built:
  - - Input: (batch, 30 days, 11 features)
  - - 3 TemporalBlocks with dilations [1, 2, 4]
  - - Channel progression: 32 → 64 → 32 filters
  - - Output: Single gold price prediction
- Key Code Components:

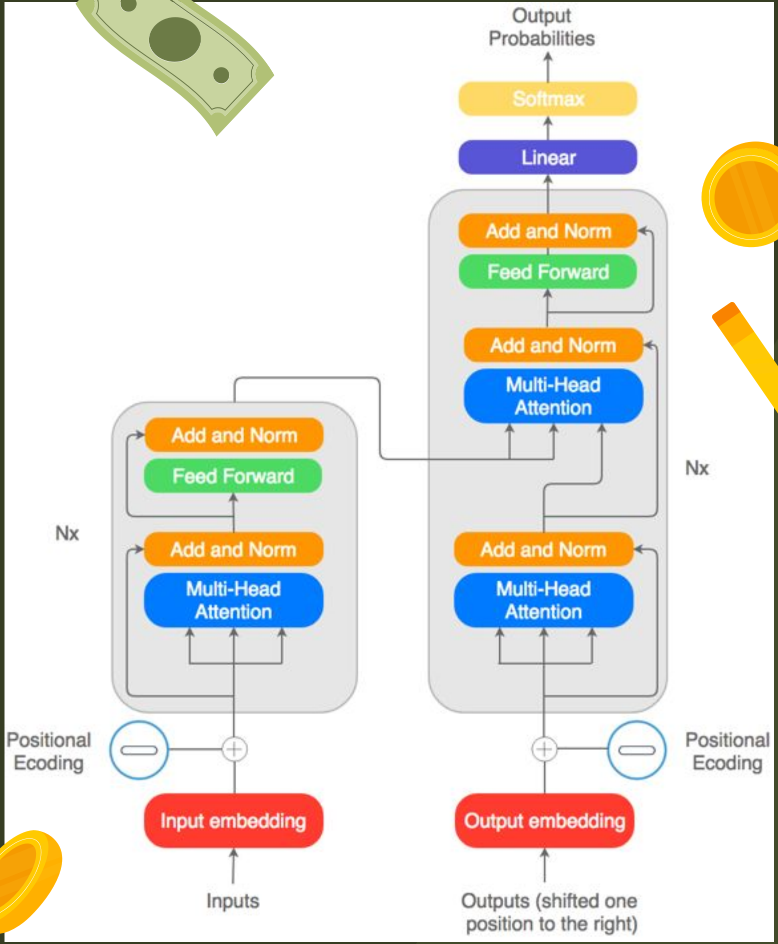
  Causal Convolution:

```
class Chomp1d(nn.Module):
    def forward(self, x):
        return x[:, :, :-self.chomp_size]  # Remove future!
```

  - Temporal Block:
    - - Conv1D → Chomp → ReLU → Dropout → Conv1D → Chomp
    - - Residual connection: output = input + processed_input
    - - Dilation pattern: 2^i (1, 2, 4)
  - Final Model:
    - - TCN backbone + Linear layer
    - - 36,641 total parameters
    - - Trained in ~56 seconds

Transformer Model

- ***Data Preparation***: TimeSeriesDataSet with encoder length 30, prediction length 1 (feature dim = 11)
- ***Core Model***: PyTorch Forecasting's TemporalFusionTransformer
  - Hidden size: 64, Attention heads: 4
  - Dropout: 0.1, Continuous size: 32
- ***Custom Model***: MVTransformerWithAttn
  - Input projection → d_model=128, n_head=8, 3 encoder layers
  - Pooling & Dense head for regression
- ***Training & Optimization***: PyTorch Lightning Trainer (epochs=10, AdamW, MSELoss)



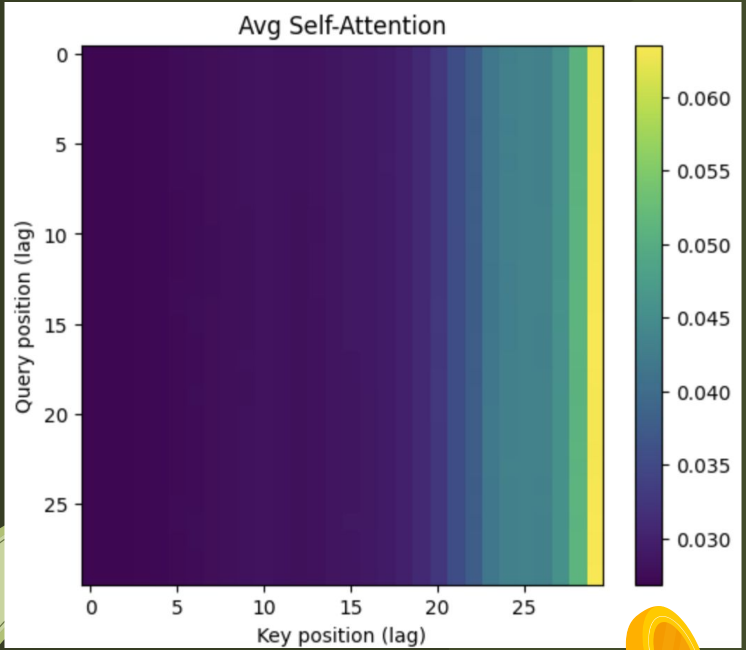
Results

Params to train = **~398k**

Time taken to train on T4 GPU = **~50 secs**

Validation RMSE = **0.148**

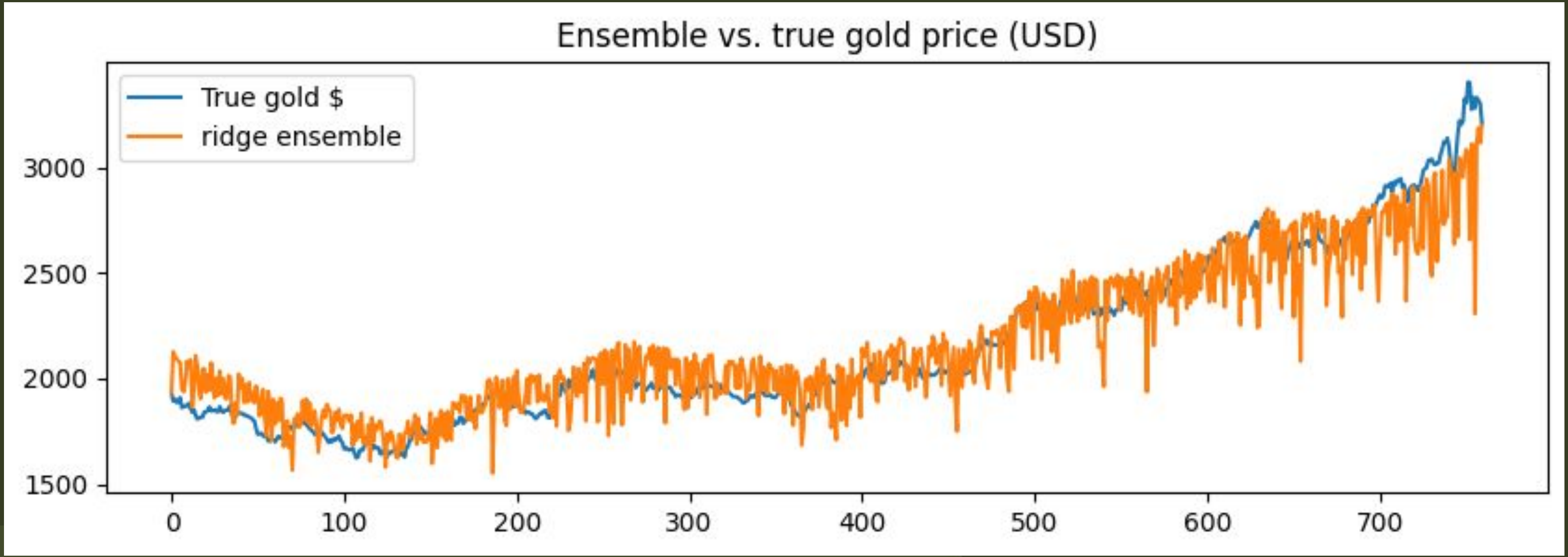
Bi-Direction accuracy = **52.7 %**



Key Observations

- **Magnitude vs. direction trade-off:**
  - Bi-GRU wins on RMSE, Transformer on DA.
  - Ridge focuses on minimizing squared error, so it leans heavily on Bi-GRU and improves RMSE but not DA.
- **TCN under-performed in comparison:**
  - RMSE > 0.8 indicates a scaling bug or label mis-alignment. It still adds slight directional signal but hurts any simple average.
- **RMSE-weighted fusion struggled:**
  - Because TCN's RMSE is an order of magnitude larger, the 1/RMSE weighting assigns ~100 % weight to Bi-GRU—identical result.

Ensemble Results



Ensemble Results

Benchmark group	Typical setup & metric	What they get	Our result
Naïve baseline(“tomorrow = today”)	RMSE on scaled prices	0.09 – 0.10	0.067
Academic papers on gold-price forecasting(LSTM / GRU with tech indicators)	RMSE on *unscaled* USD prices	110–200 USD (≈ 0.07–0.12 when scaled to our range)	≈ 150 USD (0.067 scaled)
Modern SOTA research papers (Temporal Fusion Transformer, Informer, TFT + macro/news)	Directional accuracy (DA) 55–65 %RMSE 0.05–0.06 (scaled)	58% DA 0.050 RMSE	53 % DA 0.062 RMSE
Quant-fund intraday models (thousands of covariates, alt-data)	Hit-rate 50.5–54 % on high-frequency returns; edge proven via Sharpe	55-60%	53 % DA at daily horizon
Commodity banks / research desks (ARIMA + macro judgment)	1-month MAPE ≈ 2–3 %	~65–100 USD error	~4.5 % MAPE (~150 USD)

What do our results means?

- **Why our numbers are quite respectable, despite accuracy!**
- **Tiny feature set, no news sentiment:**
    - We used only four market series + economic indicators.
    - Big financial firms add FX crosses, gold ETF flows, Fed statements, Twitter/Bloomberg sentiment, etc., to shave the extra few RMSE points.
    - They also have Billions of $'s in budget
- **Ultra-light models:**
    - Single-layer Bi-GRU & HF-Tiny weights (~0.5 M params) **trained in 5% of the time**.
    - Cutting-edge transformers run tens of layers and hours of hyper-search.

04

Q&A

THANK YOU

- Abhinav Vadhera
  - vadhera@usc.edu
- Rodrigo Lopez
  - wrlopez@usc.edu
- Vikyath Naradasi
  - naradasi@usc.edu

Do you have any questions?