Synthetic Financial Data with Probabilistic Forecasting:

A Generative Adversarial Network Approach

Mohammed Alruqimi.[1] Luca Di Persio[1]

[1]Computer Science Department
University of Verona

Women in Fintech and AI, June 2024

Table of Contents

Introduction

Synthetic financial data

- A computer-generated representation of real-world financial data.
- Unlike real-world data, which are collected from various sources such as stock exchanges, synthetic financial data is created from scratch based on predefined rules or statistical models.

Introduction

Benefits of synthetic financial data

- Data Privacy and enhanced collaboration and knowledge sharing.
- Addresses issues of limitation of data availability.
- Provides high-quality, diverse datasets for robust model training.
- Data bias reduction.
- Enables stress testing of financial systems under hypothetical scenarios.

Challenges

Challenges

- Complexity of financial markets.
- High dimensionality and correlation in financial data.

Problem

Insufficient accuracy in reflecting real-world complexities.

Literature review

Statistical Methods: early attempts employed traditional statistical methods such as (bootstrapping, Monte Carlo simulations, Stochastic Differential Equations (SDE), and time series models such as ARIMA (Autoregressive Integrated Moving Average).

Variational Autoencoders (VAEs): a deep learning technique where the encoder maps input data to a latent space, and the decoder generates new data points from this latent space.

GANs: a deep learning approach where two neural networks (generator and discriminator) are trained together to produce data that is indistinguishable from real data.

Proposed Approach

Use a conditional GAN network [MO14] to generate future crude Brent oil prices based on historical observations and relevant variable(sentiment analysis of the crude oil market). The model is built on the ForGAN model [KSDA19].

GRU-CGAN Model

The GRU-CGAN model employs a Gated Recurrent Unit (GRU) architecture for its generator and discriminator network.

Dataset

1. Historical daily observations of the crude Brent oil Price (2012-2021).
2. Daily sentimental index (2012-2021), generated using the crude BERT model.

GRU-CGAN Model - Generator (G)

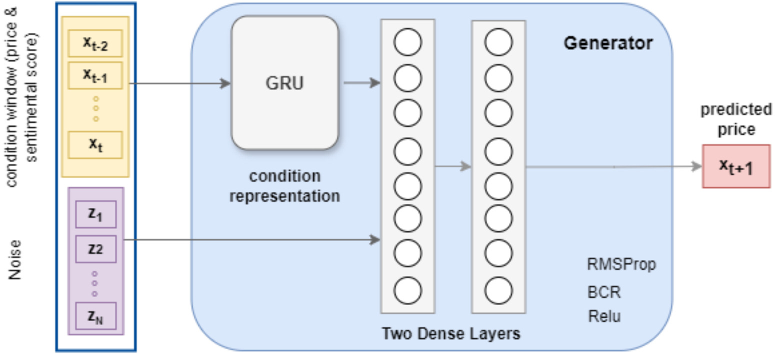


Figure: The generator network

GRU-CGAN Model - Discriminator (D)

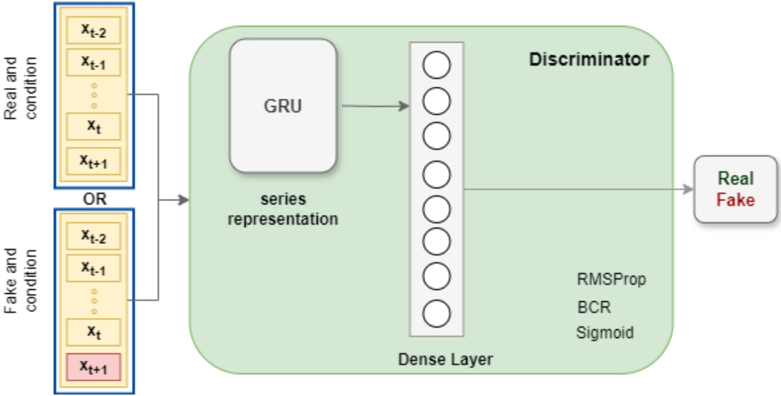


Figure: The discriminator network

The objective is to model the probability distribution of one step ahead value x_{t+1} given the historical data $c = x_0, .., x_t$, i.e. $\rho_{(xt+1|c)}$.
The Generator G and Discrimniator D are trained simultaneously in an adversarial network. The generator G learns to transform a known probability distribution ρ_z to the generators distribution ρ_G which resembles ρ_{data}. While the discriminator receives (x_{t+1}) and determines if (x_{t+1}) is real or generated by the Generator. Hence, the model function is expressed as follows:

$$\min_G \max_D V(D, G) = \mathbb{E}_{x_{t+1} \sim \rho_{\text{data}}(x_{t+1})}[\log D(x_{t+1}|c)] + \mathbb{E}_{z \sim \rho_z(z)}[\log (1 - D(G(z|c)))]$$

Experiments

- Hyperparameters and model training

  To avoid excessive complexity, we first trained the generator independently as a stand-alone model, applying the Grey Wolf Optimizer (GWO) to find near-optimal hyperparameters.

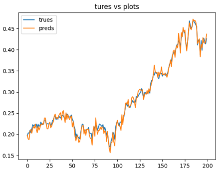
- Evaluation metrics Point-wise error metrics are used to our model with benchmark models, we report MAE, MSE, RMSE, MAPE, and MAPE.in our model with benchmark models; Additionally, we used Kullback-Leibler Divergence (KL Divergence) to measure the distribution similarity between the actual and generated data. in our model with benchmark models;

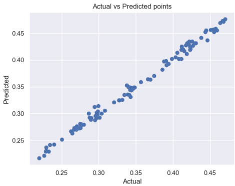
Table: Evaluation metric

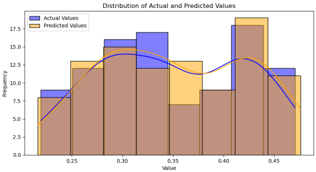
MAE	MSE	RMSE	MSPE	MAPE	KL
0.003900	0.000062	0.006481	0.000459	0.0.012710	0.000183

investigated the use of a conditional Generative Adversarial Network (GCN

Figure: Evaluations of the proposed model: Actual vs generated values for Brent dataset







Conclusion

In this work, we investigated a conditional Generative Adversarial Network (cGAN). We trained the GAN model in a supervised learning approach by incorporating sentimental scores and historical observations as conditioning inputs) for generating a synthetic time series dataset of Brent crude oil prices. By incorporating both sentimental scores and historical observations as conditioning inputs, we trained the GAN model in a supervised learning approach. The results demonstrated that the model successfully produced data with a high degree of similarity to the original dataset.
The model was employed to generate 200 data points, and the evaluation metrics indicated a strong performance. The Mean Squared Error (MSE) between the generated and original data was found to be 0.000062, while the Kullback-Leibler (KL) Divergence was calculated to be 0.000183. These metrics were assessed using normalised data.

References I

Alireza Koochali, Peter Schichtel, Andreas Dengel, and Sheraz Ahmed, *Probabilistic forecasting of sensory data with generative adversarial networks – forgan*, IEEE Access **7** (2019), 63868–63880.

Mehdi Mirza and Simon Osindero, *Conditional generative adversarial nets*, CoRR **abs/1411.1784** (2014).