Learning Embeddings of Financial Time Series Through Multitask Classification

Khoa Tran[1], Joe Wang[2], Mehdi Zouiten[3]

[2]LexisNexis Risk

Introduction

- **Introduction to representation learning and embeddings in NLP:** Representation learning, particularly in NLP, focuses on automatically discovering the representations needed for feature detection or classification from raw data. Embeddings, which are dense vector representations of words, have revolutionized the way machines interpret text by capturing contextual meanings, semantic relationships, and the nuances of language.
- **The significance of applying these concepts to financial time series:** Just like words in NLP, financial instruments can also benefit from nuanced, high-dimensional representations. Financial time series, comprised of complex patterns and non-linear relationships, can be better analyzed using embeddings, which can capture underlying market dynamics and offer more sophisticated insights into price movements and trends.
- **Novelty of the project in the context of financial data analysis:** Our project breaks new ground by leveraging the principles of representation learning from NLP to financial time series. By creating multifaceted embeddings of financial data, we provide a fresh perspective on the market, which has the potential to outperform traditional analysis and prediction models that rely on more simplistic or linear data interpretations.

Objectives

Development of representations for financial time series: The project aims to create a comprehensive framework for representing financial time series data by encoding them into embeddings. This involves transforming raw numerical data such as stock prices and financial ratios into a learned vector space where distances between vectors correspond to the similarity of the underlying financial conditions. These representations are expected to capture both the temporal dynamics of stock movements and the intrinsic characteristics of companies and sectors.

Facilitate sector classification and market regime prediction through multitask classification: The embeddings are used to serve multiple predictive purposes. First, they aid in classifying stocks into their respective industry sectors, allowing for a nuanced understanding of company group dynamics. Secondly, they help predict market regimes, such as the onset of a recession or a bull market, enabling investors to make more informed decisions. Multitask classification implies that the model simultaneously learns to perform these tasks, leveraging shared representations for improved generalization and performance on each individual task.

Data Structure

Consolidated Time Series: Our dataset aggregates monthly stock metrics with quarterly financial reports from leading U.S. companies, capturing essential temporal trends in the financial market.

Comprehensive Feature Selection: We include a suite of financial indicators like profitability and debt ratios, each serving as a dimension in our feature space for multifaceted analysis.

Data Enrichment and Integrity: Records are enriched with industry sector and regional classifications, while a forward-filling approach maintains continuity in the face of data gaps.

Timestamp	Ticker	Feature 1	...	Feature N	Sector	Region
Month1	Ticker 1	2000.00	3000.00	4000.00	Sector n 1	0
Month1	Ticker 2	3000.00	4000.00	5000.00	Sector n 3	0
.	...	...	...	...	...	0
.	...	...	...	...	...	0
Month i	Ticker 1	2000.00	3000.00	4000.00	Sector n 1	1
Month i	Ticker 2	3000.00	4000.00	5000.00	Sector n 3	1
.	...	...	...	...	...	0
.	...	...	...	...	...	0

Figure 1. Dataset Structure

Model Structure

- Our architecture is inspired by the Transformer model, which excels in processing sequential data. Key modifications include the omission of positional encodings, given our hypothesis that the self-attention mechanism inherently captures the temporal sequence of financial time series data. Additionally, we introduce tailored layers and a secondary branch to capture global financial patterns essential for market analysis.

Experimental Setup

- Our model is a transformer-based architecture tailored for the time-dependent nature of financial data. Modifications have been made to standard transformer structures to better suit the nuances of the financial domain.

Experimental Setup

- Multitask learning is achieved through a classification that discerns between industry sectors and market regimes.
- The CrossEntropy Loss (L_{CE}) captures sector classification accuracy:

$$L_{CE} = -\frac{1}{N}\sum y_{ic}\log(p_{ic})$$

- Center Loss (L_{Center}) ensures feature clusters are tight and distinct:

$$L_{Center} = \frac{1}{2}\sum \|x_i - c_{y_i}\|^2$$

- Focal Loss (L_{Focal}) addresses imbalance by focusing on challenging cases:

$$L_{Focal} = -\alpha(1-p)^{\gamma}\log(p)$$

- The Total Loss is the aggregate of the three, balanced by λ:

$$\text{Total Loss} = L_{CE} + L_{Focal} + \lambda L_{Center}$$

Results

- To assess model performance, we measured precision, recall, F1-score, and ROC-AUC score. These metrics provide a comprehensive view of our model's accuracy and ability to distinguish between classes.
- We compared our model against baselines to demonstrate improvements, especially noting higher F1-scores in sector classification and a superior ROC-AUC score in market regime prediction.

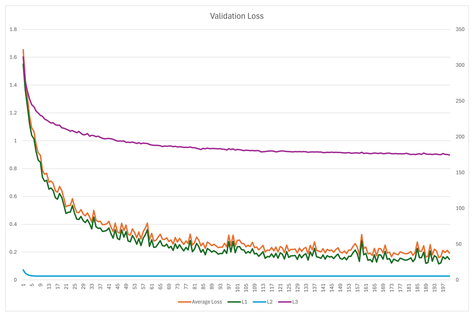

Figure 2. Validation loss plot showing the average loss and individual loss curves L1, L2, and L3 over the epochs

Results

- Through the application of t-SNE, we project our high-dimensional financial embeddings into 2D and 3D spaces, allowing for an intuitive visualization of the data's internal structure. These visualizations showcase clear sector-specific clusters and subtle distinctions between market regimes, indicating the model's adeptness at capturing key financial dynamics and relationships. This not only confirms the effectiveness of our embeddings in representing complex financial time series but also highlights their potential to uncover insights into market behaviors and sector interrelations.

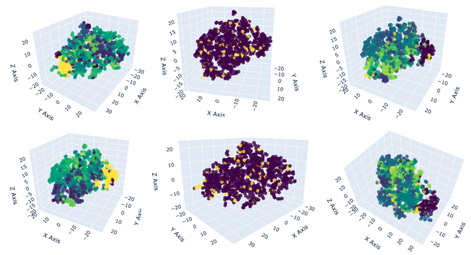

Figure 3. Visualization of t-SNE clusters. From left to right is sector, regime, joint labels. First row is train set, second row is test set

- We enhanced our analysis by integrating K-nearest neighbors (KNN) with our embeddings for sector classification and recession prediction. Utilizing the L2 distance metric, we systematically explored various k values to fine-tune accuracy. While sector identification showed promising accuracy, recession prediction faced challenges, likely due to data imbalance and the inherent limitations of KNN in high-dimensional spaces. This highlights the nuanced complexity of financial time series data and the need for tailored approaches to different classification tasks.

Conclusion and Limitations

In this study, we developed embeddings for financial time series, integrating returns and fundamental data to capture industry sectors and market regimes. Our Transformer-based model, enhanced with metric learning and computer vision techniques, showed promising sector classification and recession forecasting capabilities. Performance metrics on the test set and t-SNE visualizations confirmed the model's effective learning and generalization abilities.

However, challenges like label imbalance and the inherent limitations of KNN in high-dimensional spaces were encountered, affecting recession prediction accuracy. Future work will aim at refining the model through hyperparameter optimization, exploring alternative algorithms, and addressing data imbalance to enhance performance further.

Overall, our findings highlight the potential of embedding techniques in financial data analysis, setting a foundation for future advancements in this field.

References

1. Sokolov, Alik, et al. "Neural Embeddings of Financial Time-Series Data." The Journal of Financial Data Science 2.4 (2020): 33-43.
2. Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017)
3. Sun, Yanfeng, et al. "A financial embedded vector model and its applications to time series forecasting." International Journal of Computers Communications Control 13.5 (2018): 881-894.
4. Dolphin, Rian, Barry Smyth, and Ruihai Dong. "Stock Embeddings: Representation Learning for Financial Time Series." Engineering Proceedings 39.1 (2023): 30.