University of Warsaw
Faculty of Mathematics, Informatics and Mechanics

Damian Dąbrowski
Student no. 439954

Heorhii Lopatin
Student no. 456366

Ivan Gechu
Student no. 439665

Krzysztof Szostek
Student no. 440011

Large language models for forecasting market's behaviour

Master's thesis
in COMPUTER SCIENCE

Supervisor:
dr Piotr Hofman
Instytut Informatyki

Warsaw, May 2024

Abstract

This thesis concerns research into the use of machine learning and large language models in market analysis, focusing on market predictions.

Keywords

machine learning, large language models, time series forecasting, market prices

Thesis domain (Socrates-Erasmus subject area codes)

11.4 Sztuczna inteligencja

Subject classification

D. Software
D.127. Blabalgorithms
D.127.6. Numerical blabalysis

Tytuł pracy w języku polskim

Duże modele językowe w przewidywaniu giełdy

Contents

Chapter 1

Introduction

In the world of stock markets a major problem is the apparent incalculability of the complex network of factors e.g. how stock prices of one company affect those of another etc. As the environment of stock markets becomes more and more complex, the ability to analise and confidently predict its future becomes of crucial importance for traders, investors and researchers.

With the recent advent of generative AI and the demonstrable power of Large Language Models a question arises of how these can be used to accurately analise and predict time series market prices in different environments. Therefore, this thesis presents our work on the subject.

First, we look at what work has already been done in the field of LLM time series prediction, in particular what techniques of fine-tuning and input data transformation were used. Then we look at how different, smaller machine learning models deal with time series prediction.

Subsequently, we discuss our methodology; different aplied methods and techniques of input reprogramming, use of prompts and context, and LLM fine-tuning. Next, we present the results we have achieved on the chosen datasets (and compare them to some other known solutions).

Finally, we speculate on the significance of our work, its potential applications in forecasting price time-series.

Chapter 2

Related work

Chapter 3

Other models

Here we present our results from trying to use the following models to extrapolate a time series. Classification models output in binary categories: increase or decrease in value, and are therefore less precise.

3.1. Random forest

Random forests were tested only on the simple sales data. The following options were tried
- num_lags: [1, 5, 10, 13, 25, 40, 50]
- n_estimators: [20, 50, 100]
- max_features: [2, 4, 8]
- criterion: ["gini", "entropy", "log_loss"]

The accuracy of the models ranged from 0.749 to 0.827. The main influence seems to be the lag number - the optimal being around 10. Then slightly better are those with max_features = 4, and number of trees ≥ 50. The criterion doesn't seem to play a significant role.

3.2. Support vector machine

For the support vector machine there were overall nine models tried:

/noindent - three kernels: ‘rbf‘, ‘poly‘, ‘sigmoid‘
- one gamma: ‘scale‘
- three C values: ‘0.1‘, ‘1.0‘, ‘10.0‘

The ‘poly‘ kernel worked much better than both ‘rbf‘ and ‘sigmoid‘, which both worked equally badly. Overall, though, the statistics for every model were terrible. The value of ‘C‘ has had very little impact and only on ‘rbf‘ kernel.

3.3. Multi-layer perceptron

3.4. Convolutional neural network

3.5. Residual neural network

Chapter 4

Methodology

Chapter 5

Main results

Chapter 6

Forecasting applications

Chapter 7

Conclusion

Appendix A

Visualisation

Bibliography

[Bea65] Juliusz Beaman, *Morbidity of the Joll function*, Mathematica Absurdica, 117 (1965) 338–9.

[Blar16] Elizjusz Blarbarucki, *O pewnych aspektach pewnych aspektów*, Astrolog Polski, Zeszyt 16, Warszawa 1916.