Transcription for diploma presentation

Andrey Grishin

March 28, 2023

1 Slide (Start)

Hello, my name is Andrey Grishin and today I am going to tell you about my research connected with exploring the performance of Econometric and Neural Network models on empirical data of Chinese and US markets.

2 Slide (About plan)

First of all it is crucial to mention the plan of the presentation. We will start with the introductory part that contains motivation, targets and tasks that should be completed to achieve initial goals. Then comes pre-experiment part with the foundation of my research, data analysis and insights and finally we will discuss the methods that were used to prove further established hypothesis. In conclusion, we will see the achieved statistic results up to which the question — whether the initial hypothesis was approved or rejected — was answered. So, let's start.

3 Slide (Briefly about today)

In today's world there is a plenty of various words in daily routine of each person. However, the most common one is "money". It means that if we assume rationality of the individual — here rationality means desire of a person to do everything that he/she wants and nothing that he/she does not want — we get the aspirations to satisfy personal wishes. Hence, it can be interpreted as maximization of utility or — in terms of money — income.

From this point the necessity of development of math models for forecasting required indicators is obvious. Due to this research, I am talking about stock market and prediction of stock returns and — as we will see further — prices.

Initially there were developed many statistic and econometric models, but in case of the Internet development the empirical data amount grew too fast. That is why the new sphere of science, called Data Science, was developed. In contrast, nowadays there is a huge number of math models, that can be used to examine data.

So, this fact is the catalyst of competition between econometric and Machine Learning approaches to gathered data analysis. And as a subset of Machine Learning, Deep Learning is also competing with econometric and statistic approaches.

4 Slide (Motivation — why)

According to the motivation sphere here we have 2 main branches: theoretical and practical. In case of theoretical one — figuring out the most effective forecasting model (due to the set of

them analyzed in my research) for developed and developing markets allows not to use other models and improve the only selected one. So the problem of model choice will be solved. And in practical case it will become much easier and, hence, much quicker to answer the "Buy, hold or sell" question. So the computer algorithms will be able to make much more reliable decisions that will help people to make money. As a result investors, traders and stock players are happy.

5 Slide (Targets — what we want to achieve)

Targets setting is a very important part of any research, that is why here are only three things, going in the following order. The main target of this work is to help traders to make more accurate decisions on "Buy, hold or sell" questions. Then, applying further mentioned algorithms user will be able to make more secure deals, hence profitable 'cause it will be possible to make Long Term forecasts. Finally, due to the number of different indicators, indexes, figures, ratios it is very difficult for not qualified user to master trading and start make money, so figuring out the best algorithm will allow this person to feel much more self-confident in investing and stop being scared of stock market.

6 Slide (Tasks)

To be more precise, in this research I provide sequential models' comparison based on empirical dataset of prices and returns of US (15) and Chinese (15) companies. US was taken as a developed economy, where as China (Shangahi not Hong-kong) as developing one.

7 Slide (Hypothesis — what was assumed)

To say more about foundation of my research it is impossible not to mention 3 main hypothesis. Initially in 1970 Eugene F. Fama published his article about market efficiency, saying that it is impossible to predict volatility or price of any asset because it already includes information about everything in itself . The only thing that can help to forecast is news. After this in 2006 Benoit B. Mandelbrot published another work, telling that market processes have long memory — it means that there are persistent and anti-persistent processes in market reality. Next, due to Martin Sewell's article, published in 2011, we have that hypothesis of Market Efficiency is "The best thing for today, but is not true", that is why it is worth trying to develop models that will help in forecasting process. This conclusion was made in case of analysis of all previously known materials about market efficiency. Finally I would like to formalize my own hypothesis confirmation or rejection of which will be discussed further. I am trying to check if neural network approach for stock returns prediction is the best one for developed and developing markets.

8 Slide (More about data)

In this research I use open prices and its returns of 15 american and 15 chinese companies. United States companies were chosen as an example of developed market and chinese — as developing one. Duration of each time series is from companys' IPO date till December 13 of 2022, because after this date international markets are experiencing unstable behavior caused by political decisions.

9 Slide (Data insights from visual analysis)

For better understanding data we are working with, let's look at this example. On the left — we have Coca Cola and on the right — Kweichow Moutai — on the biggest (in terms of capitalization) Chinese alcohol drinks producer. Looking at this graphs we notice **two** things. The first one is that both of upper time series (prices) have increasing tendency with higher volatility through time. The second one is that looking at the lower time series (returns) we see that volatility of US company open price is in the interval from -10% to 10% with more confidence than the same indicator for Chinese company. These facts tell us that developed and developing markets differ from each other for example by volatility behavior, hence with higher probability it is required to use different models to describe them. And also, looking at returns figures, there is a slight thought that it is much more difficult to forecast it, than prices. And the change of returns to prices analysis do not cause too much damage because it is very easy to compute returns, if we know prices. However, it is just an assumption with no proves yet. To provide them, let's look at the time-frequency domain of returns and prices using Wavelet transform principle known from quantum-physics. Our goal is to know if there are any dominant frequencies in theses "signals" that can be used in forecasting purposes.

10 Slide (Scalogram for Coca Cola)

Due to the scalogram of Coca Cola prices and returns we can find out that prices do not contain frequency information where as returns contains a lot of frequency significant information. Here under "frequency" evidence I mean repetitive patterns of graph form, that can be used in modeling. So, we have a problem. Is it worth to analyze just returns but not prices?

11 Slide (Scalogram for Kweichow Moutai)

No! Looking at the left part of the slide we can see scalogram of Kweichow Moutai prices. Here there is no need to say that they obtain extremely important frequency insights. So, as long as returns have similar but less significant things, it is better not to skip prices from analysis.

12 Slide (Applied methods)

In my research I used 15 models 8 of them can be classified as statistic and econometric approaches and other 7 are network approaches. I do not use word "neural" here, because for 2 models of them (WN and MSSA + WN) this word is not true, as they are based on slightly other working principle.

In the first set of 8 models there are EWMA, ARIMA, ARIMA + GARCH, ARIMA + FIGARCH, ARFIMA (Stands for Fractionally Integrated Autoregressive Moving Average), ARFIMA + GARCH, ARFIMA + FIGARCH, SSA. All of them are associated with empirical data, however they do not require too much of it for fitting, hence there is less confidence of accuracy in contrast with network approaches which require a lot of preprocessed data for good fitness of parameters.

According to the second set of 7 models we have: MLP, MLP + MSSA, MLP + EWMA, RNN, RNN + MSSA, WN, WN + MSSA. What does it mean? MLP stands for Multi-layer Perceptron firstly invented by Frank Rosenblatt in 1961 and further this concept was improved to such things like today neural networks — Convolution and Recurrent ones. The last one is also used in this research. MSSA stands for Multistage Singular Spectrum Analysis invented by Kuang, Wang, Lai, Ling in 2020 for no-hyperparameters signal denoising process. It is

the same thing as SSA (Singular Spectrum Analysis). WN stands for Wavelet Networks invented by Zhang and Albert Benveniste in 1992 and describer by Antonios Alexandridis and Achilleas Zapranis in 2014. These networks are used as another attempt to describe signals with frequency patterns. All in all it does not matter what exactly these models mean, because these models are different approaches to the one thing, that are connected by the definition of neural networks. So, the main thing that should be done is comparison of these models on empirical data. Therefore it is vitally important to have metrics function, that will numerically measure each model's forecasting ability. This metrics is WAPE (Weighted Average Percentage Error), that can be interpreted as percentage deviation of model's predictions from ground truth.

13 Slide (Result for prices)

According to the collected results, we can see that, obviously "yes" neural approach is extremely good at price forecasting, as the minima error is 1.06% for developing market and 0.93% for developed market on testing set, comparing with average 120% for econometric models. However, it is impossible not to notice that EWMA and SSA — statistic approaches — are really good too. So, to summarize information from this chart, we see that statistic and econometric models such as EWMA and SSA are the best, however the global winner is MSSA + MLP model, that achieves 1.06% for developing and 0.93% for developed markets. So, that is about prices, but now let's look at returns' results.

14 Slide (Result for returns)

On the one hand, returns' result is much-much poor either for econometric models or neural networks. On the other hand, we have an interesting fact that, RNN + MSSA beats all other models totally. So, neural network approach is the best for returns forecasting.

15 Slide (Discussion of result)

Finally, let's gather out all new important insights, that were learned from this research. 1) It is better to forecast prices, because prediction accuracy is much higher than in returns foreseeing. 2) MLP + MSSA model is the best for prices forecasting. 3) MSSA + RNN is the best for returns forecasting. 4) Econometric models are bad for prices and returns forecasting. 5) Econometric models are less accurate — in terms of prices — for developing market. 6) MSSA algorithm of denoising has positive influence on forecasting accuracy, but Wavelet Network is an exception. 7) WN and MSSA + WN training process took up to 7 days of non-stop working. 8) EWMA is the fastest for modeling and can be called a simple RNN. 9) Boosting approaches, that were used in combination of models like AR(F)IMA + (FI)GARCH, is a bad thing, so it is assumed to be much better in case of minimization of loss function simultaneously for combined model and not training the second one on the residuals of the first. 10) ARFIMA is better than ARIMA, hence market fractallity exists.

16 Slide (Conclusion)

That is all. Thank you for attention!