Automated prediction of stock markets returns, based on wide spectrum of financial statements

Author: Krasimir Krastev

Abstract: Historical stock markets data receive attention by researchers since the early days of modern econometrics. Many believe that historical data contains important models, which can be used for tracking of future movements in stock prices. But in many cases, this information offer only limited (and very often biased) overview of some of the indicators for public companies and offer very little context for the long term internal processes. For this reason, models based solely on historical returns in general performs poorly and can be useful only for short term modeling of returns. After entering in the digital era, significant part of the management processes of companies became computer-driven. This made access to internal accounting data much easier and offered unique possibility to include various indicators from companies balance sheets for improvement of existing forecasting models. In this paper will be reviewed innovative ML-based system for generation of predicting models, proposed in recent paper by Bogdanova et. al. (2021), based on combined input from historic price data and set of variables from accounting reports.

Keywords: Machine Learning; Stock Markets; Accounting Data; LASSO

JEL: G17

1. INTRODUCTION

Emotions of investors (i.e. greed, fear or panic) plays a major role in volatility of international stock markets (e.g. Lee et. al. 2002, Baker et. al. 2012, Koppel et. al. 2021) and can be differentiated as separate risk factor of its own. The effect of this factor on the movements of stock prices is catalyzed by dramatic events (such a major financial collapses, unexpected crises or global speculative bubbles), when rational thinking of investors is suppressed by psychological biases. One of the extreme examples is the bursting of Dot-Com bubble in 2002, which caused market panic through massive sell-off of tech companies stocks and driving down the price of their shares to unsustainable levels. The losses of financial markets caused by this event was around $5 trillions and major stock exchanges took heavy blow (NASDAQ index lost nearly 78% of its value). Another example is the COVID-19 crisis, where uncertainty and fear around the pandemic caused over-pessimistic investor sentiments (Dash et. al, 2022). It is more likely for less experienced investors to be affected by emotional bias (and to accumulate higher losses), due to overreaction caused by stressful events.

Emotions will always persist in financial markets, as human factor cannot be fully isolated from decision-making process. In order to minimize its negative effects, it is necessary to incorporate extensive AI assistance in all steps of creation and execution of investment strategies. Automated evaluation of the impact different factors can have on the market, can help to recognize and negate cases of induced emotional bias. Critical point of AI-based approach is to supply sufficient amount of data for identification of ML-based models. The quality of those models will determine the effectiveness of the AI. If the data is heavily biased by emotions, ML will struggle to recognize important patterns and will perform poorly. One of the possible improvements is to enrich existing historical data with additional non-biased information. However additional problem arise – how to identify suitable information, which is not sentiment-augmented and hold predictive power for financial markets.

The Efficient Market Hypothesis derived from the research work of Fama (1970; 1991) is important milestone in theory for the impact of the information on financial markets. According to this hypothesis, the efficiency refers to how well the markets prices reflect available information. Fama defines three different levels of efficiency: strong, semi-strong and weak. In strong-form all available information (public or private) is accounted for stock prices. Weak-form assumes that all future securities prices are random and not influenced by past events. Semi-strong form implies only public available information in current stock prices.

Weak efficiency form is most studied and significant amount of evidences are provided against its validity (i.e. Bogdanova et. al 2021, Bogdanova 2021). For evaluation of the possible relations between historic data and future movement of stock prices, suggested by remaining forms of efficiency, multiple methodologies are possible.

Traditional fundamental approach for generation of prediction models integrates financial performance of historical data for securities or market indexes. This methodology can be useful for extraction of behavioral patterns, based on variety of different variables, identification of regular economic cycles or unexpected, exogenous world events. Inherited disadvantage of those models, which significantly limits their usability is that they capture very limited abstraction of complex real-world dynamical systems, involving uncounted number of factors. Impact of those factors is not static as well and can evolve dramatically over time. Those two properties are catalyst for incremental degradation of predicting power in future horizon.

In contrast to the fundamental analysis, which concentrate on financial performance of securities, technical analysis (see Kirkpatrick et. al. 2006) incorporates additional market information (i.e. various volume/price transformations, money flows, momentum) in order to forecast directions of future movement in the prices. This type of analysis tries to identify trends or patterns in historical data, which may align with current financial and economic conditions. However this methodology is most useful in short-term forecasts for assets with frequently fluctuating price movements, such a commodities.

Technical analysis can be further enhanced for securities by incorporation of companies financial statements. The predicting power of accounting data for forecasting of stock returns is still considered controversial, but this topic is drawing increased attention by

researchers in last decades (i.e. Kirkpatrick et. al. 2006, Ou et. al. 1989, Noma 2010, Nissim 2022). Unfortunately the data from financial reports cannot be utilized directly into machine-learning models for stock markets returns. Main reason for this setback is that integration of data from financial statements into regression models is very difficult due to limited amount of reports, compared to much bigger count of metrics used as possible predictors. Another reason is related to the lack of uniform format of financial reports and the scattering of the data across different tables all around the reports. Finally financial reports are published on quarterly basis with some delay after the end of fiscal period (very often 45-60 days).

Some studies identified correlation between abnormal market returns and specific measure in accounting data. Basu (1977) noted during investigation of the relation between performance of securities against their earning-price ratios (EPR), that portfolios with low EPRs on average earned higher absolute and risk-adjusted rates than high EPR portfolios. He made the conclusion, that released public information for companies do not impact the stock market prices instantaneously and EPR can be used as early indicator for future movements in stocks. Ball et. al. (1968) examined accounting data and founded that net income receive particular interests from investors and reflect stock prices. La Porta (1996) was trying to explain the reasons for high stock returns and evaluated the possibility for mispricing by investors. He concluded that financial analysts are immensely affected by the grow rate in their forecasts.

Never the last, the scientific advancement in the field of AI for financial markets are still not widely integrated by the industry. This mistrust from the side of professional investors is caused by inherited complexity of the AI systems and the gap between reasoning of AI systems and humans. It won't be possible to convince investors to adopt AI and allow it to alter their own judgement, if they threat it as a black box. AI should be able to reason with investors and explain its decisions using their professional terminologies and practices. This will allow smooth transition to augmented trading AI, where people and machines working together to compensate for their individual deficiencies.

This paper aims to provide validation evidences for the integration of AI in financial sector and to highlight AI as future integral part of investment decision-making process. Second goal can be crucial as balancing factor agains the sentiments of investors in critical situations and to contribute for more sustainable and rational global markets. With the help of the reviewed literature is designed AI-based system for forecasting of stock-price fluctuations of specified publicly traded companies.

Most distinguishing features of this system are significant speed of reports generation (inherited by the performance of the internal regression model) and the clear interpretation of obtained results. One of the major improvements to this system, proposed in this research is the capability to train models based on combination of companies sharing common features. We assume, that this strategy will will help to identify most important data in financial statements. Another important modification is the integration of a mechanism for recovery of missing data in financial statements, which increase with significant percentage the scarce amount of available financial information.

2. METHODOLOGY

The foundation of this paper is based on the research work of Bogdanova et. al. (2021) and Bogdanova (2021), which discuss the implementation of innovative two-stage ML-based framework to support investors. In the first stage the proposed system aggregates historical stock-markets data with information from financial reports and other relevant information into integrated data-frame with year-quarters scale. Generated data-frame is used as input of Stage 2, where after multiple steps consisting of automated selection and model fit of most important features using LASSO method for logistic regression (see Tibshirani, 1996), the system is able to compute the probability for future movements of stock-prices. In sequential modules of Stage 2 additional validation on selected features is performed using rolling-sample and out-of-sample methods.

In proposed variation of the original solution, the initial two-stage design is retained, but significant changes are applied to some elements of the core logic. The end goal of those modifications is to improve resilience of the preprocessors against inconsistent/missing input data, to automate different steps inside each stage and to increase predicting power of the model.

Described solution is developed in R and used libraries for provisioning of input data overlaps with described by Bogdanova (2021). Those are "*quantmod*" (sourcing historical stock market data from Yahoo! Finance) and "*simfinR*" (integration of SimFin platform for R, source of aggregated historical fundamental financial data).

Module responsible for provisioning of accounting data in Stage 1 is improved by provisioning of catalogue with all available companies, grouped by industry sector. Generated catalogue of available companies enable execution of the ML engine against different subsets of companies, based on specified parameters. This new feature provide functionality to train prediction models, depending on required scenario (i.e best/worst market capitalization, revenue etc.). Second important modification is integration of interpolation strategy for statements with missing reports. Interpolation of missing accounting data increase with up to 30% the amount of available metrics in financial reports for some companies and helps to fix cases with missing reports for limited subset of quarters. Selected conservative settings are allowing up to 10% of all entries to be interpolated. This guarantees that low level of bias is introduced into final models. Financial statements for companies with more than 10% missing quarterly reports, are considered corrupted and are excluded from the analysis. Selected method of interpolation is cubic spline, because it is offering better smoothness and higher approximation than linear interpolation.

The output of Stage 1 defined by Bogdanova (2021) is aggregated data frame for each company, containing extracted features from accounting data for all available quarters, which is later feed to Stage 2 of the ML-engine. A binary response variable (with values Up/Down) is defined for each quarter, which show the movement of the averaged price of the stocks in comparison with the previous period.

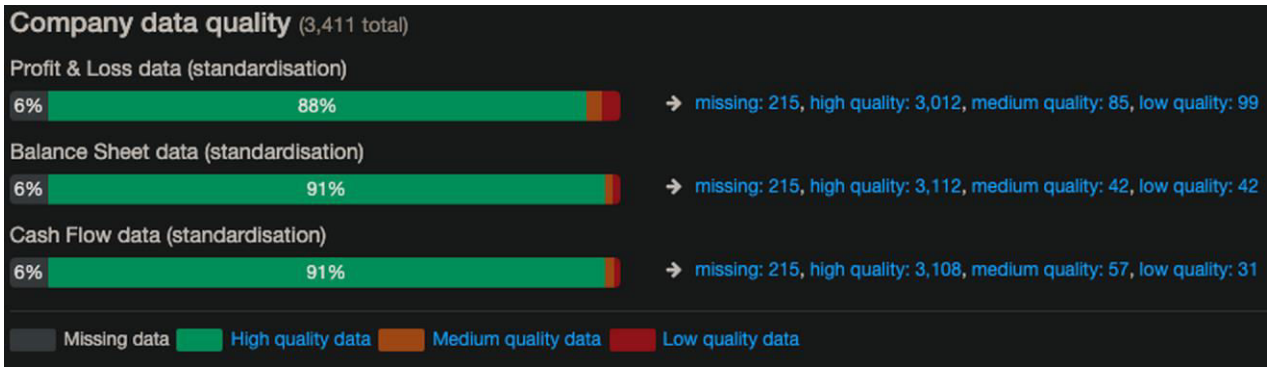
Specified number of rows from each the data-frame are reserved for cross-validation, in order to evaluate predicting power of each model.

3. EMPIRICAL DEMONSTRATION

This section demonstrates the results from practical implementation of the described design and possible future integrations inside systems for automated trading. Predicting power of models for large number of companies in different industry sector will be evaluated, in order to diversificate the research and identify metrics, which may be useful for investors.

SimFin platform provide financial data for significant amount of different companies (around 3450 companies with 630k uploaded reports, state 12/2022). Here should be noted that those companies are listed across different stock exchanges and parts of the meta-information for some of the companies is not available. Another important point to consider. is the quality of the financial data itself.

Fig. 1 SimFin - Quality for accumulated financial company data



Source : https://simfin.com/contribute/data

As shown on Fig.1 SimFin datastore contains inconsistent/missing data, which may impact the work of the ML-system. As already noted in previous section, in Stage 1 of the system is designed with some tolerance against absent entries in financial statements, which should be able to interpolate the gaps for most of the affected companies. Observation period is set between Q1/2010 – Q2/2022, while period of Q1/2021-Q2/2022 is reserved for cross-validation and is excluded from training data feed to the model.

From all of the available companies in SimFin, a catalogue containing 2702 companies grouped in 74 industry sectors in generated and on average 90 financial metrics are tracked for each company (with 239 unique metrics identified across all financial statements). Remaining companies excluded from the catalogue are not assigned to any industry sector and for this reason are excluded from the study. This catalogue serve as primary index for the ML-engine. It is useful for compilation of queries, based on metric values in financial reports or various meta-data. This advanced functionality could be extremely useful for investors, as they will be able filter-out companies from their reports, in case patterns of artificial alternation in specified financial metrics or other intentional manipulation of financial data are detected. In this study the focus is shifted agains companies with biggest market capitalization, because they traditionally receive most of the trust of investors (especially low experienced). Three companies with biggest market capitalization are selected as representatives of each sector, this altering total of 222 companies.

After a model for a company is trained with available financial statements, its predictions for the period Q1/2021-Q2/2022 is evaluated against the actual movements of

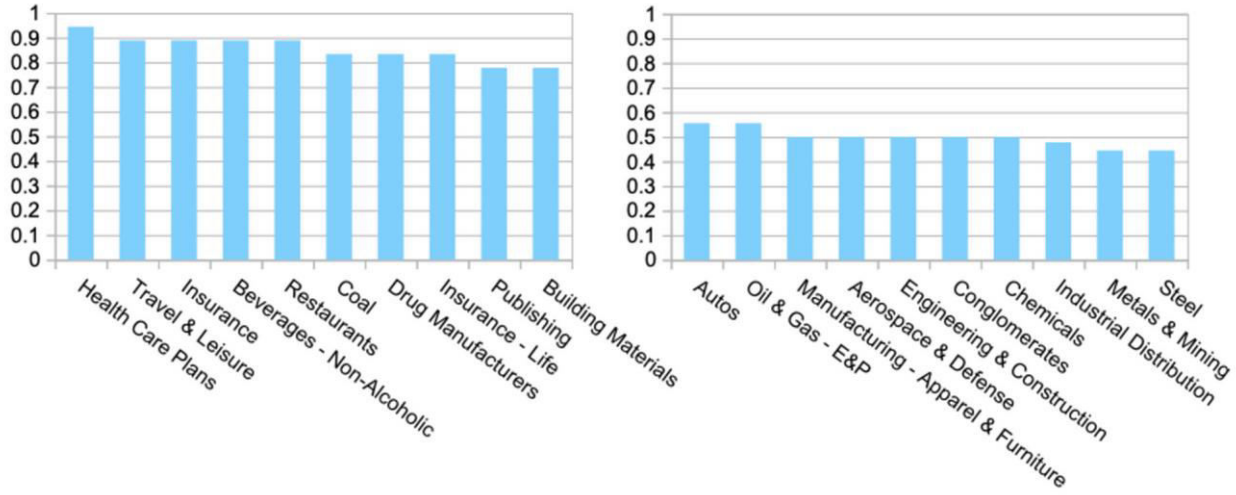
the market. Aggregated score for each model is generated as weighted sum of the probabilities for each of the quarters in test period:

$$Score = \frac{\sum Pr(Y=Y_{test}, x_q)}{N_q}$$, where x_q - quarterly financial features

N_q - total count of available quarters

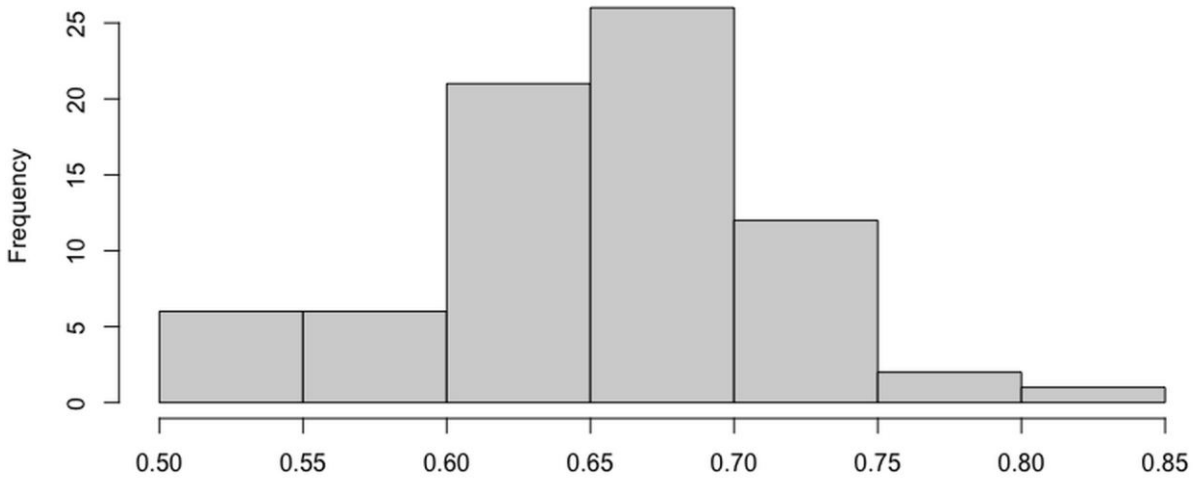
Fig. 2 shows mean value of the performance of the predicting model, grouped by industry. In both sections of the diagram are presented ten best(left) and ten worst(right)

Fig. 2 Distribution of model scores by industry (top 3 companies by capitalisation)



performing industries ordered by the predicting power of they models. Interesting tendency can be observed, if the industries are differentiated based on the type of their output. Worst prediction models was generated for industries in manufacturing sector, while those in

Fig. 3 Distribution of model scores by industry (top 3 companies by market cap)



service sector in general have 15-20% higher scores. Main reason for this difference in evaluation can be traced in more complex dynamics, involved in specifics of manufacturing industries (heavier effect of momentum due to internal inventories, more external factors such as dynamic prices of recourses and ready production).

As shown in Fig. 3, predictive power of the models, based exclusively on financial reports is situated well over 0.50 (corresponds to achieved 50% correct estimates for the future movements of stock prices), with mean value for the whole set of 72 companies at 0.661. Estimated average predicting power can be interpreted as a trace for possible correlation between financial data and the future evaluation of companies on stock markets. This contradicts to the main characteristics of the weak form of EMH discussed earlier and can contribute as additional evidence for the adaptation of remaining forms. It also highlights, that accounting information should be regarded as important source of high-quality ML data for prediction models on stock markets returns.

Another interesting point is the impact evaluation of the available financial metrics on the predicting model. For selected set of companies, 185 distinct features was identified with overall weight of all coefficients in predicting models being 127.34. Using probability density function, 20 most significant features with aggregated weight of 87.83 (corresponding to 69% of all weights) was extracted (as demonstrated in Fig. 4).

It is is important to note, that the set of predictors is not static and may alter, depending on the predefined filter during companies selection phase. This features may be useful for implementation of tailored trading strategies for any custom scenario. Additionally traders will be able to dynamically exclude/include predictors, in order to perform different simulations on the models.

Fig. 4 Most significant predictors for all sectors (top 3 companies by market cap)

income_tax_expense_benefit_net	8.9965	operating_expenses	4.4034
non_operating_income_loss	8.1212	selling_general_administrative	4.2580
net_debt_ebit	6.5088	current_ratio	3.8092
free_cash_flow_to_net_income	6.0396	pretax_income_loss	2.8823
dividend_payout_ratio	5.2229	piotroski_f_score	2.7019
other_operating_expenses	5.2196	interest_expense_net	2.5284
other_non_operating_income_loss	5.2118	cash_cash_equivalents_short_term_investments	2.1134
depreciation_amortization	4.8805	interest_income	2.0602
gross_profit	4.6129	other_short_term_assets	1.9960
cost_of_revenue	4.4955	equity_per_share	1.7682

4. DISCUSSION

Integration of AI is inevitable step in development of human civilization. In the near future nearly every human activity will have some form of integration with machine intelligence. This synergy between humans and machines will dramatically improve the performance of any manual labor, decrease the cost for the business, and will greatly decrease the negative effect of human-made errors on business processes. So far in human history every form of automatization have faced strong opposition, but it resulted in overall improvement of the prosperity in the society.

In the field of stock trading, AI can be very beneficial for humans, as it can help them to efficiently identify important patterns, in otherwise overwhelming amount of available information. In such a dynamic environment, where the speed of taking decisions and situational awareness are most important factors, presented methodology can provide

valuable and time-efficient toolset for investors, while omitting them the significant overhead related to accounting reports. Even if reviewed system is accepted into service by traders only with informative functions, it still can be very useful for justification of decisions or for identification of new opportunities.

The evaluated percentage of correct estimates are strong evidence, that prediction models build on financial statements can be very useful as supplementary tool for long-term performance determination of publicly traded companies. Another intriguing starting point for future research is the identified difference in prediction accuracy for models of companies in different industry sectors. This phenomenon could be related to the various impact of financial parameters on their respective business processes, but more detailed analyses are necessary to better understands the results.

ACKNOWLEDGMENT

The presentation and dissemination of these research results is supported in part by Sofia University Science Fund Project 80-10-159/23.06.2022 "Development of AI based algorithms for financial stability and sustainable development".

References

1. Bogdanova, B. and Stancheva-Todorova, E., 2021. "ML-based predictive modeling of stock market returns", AIP Conference Proceedings Vol: 2333, https://doi.org/10.1063/5.0042805
2. Bogdanova, B., 2021. "Applied AI in Support of Investment Decision Making"
3. Roberts, H., 1959. "Stock-markets "Patterns" and Financial Analysis: Methodological suggestions", The Journal of Finance, https://doi.org/10.2307/2976094
4. Fama, E., 1991. "Efficient capital markets: II", Journal of Finance, 46(5), pp. 1575-1617, https://doi.org/10.1111/j.1540-6261.1991.tb04636.x
5. Fama, E., 1970. Efficient capital markets: a review of theory and empirical work. Journal of Finance, 25(2), pp. 383-417, https://doi.org/10.2307/2325486
6. Basu, S., 1977. "Investment Performance of Common Stocks in Relation to Their Price-Earnings Rations: A test of Efficient Market Hypothesis.", The Journal of Finance, 32(3), pp. 663-682, https://doi.org/10.1111/j.1540-6261.1977.tb01979.x
7. Kirkpatrick, Charles D. and Dahlquist, Julie R., 2006). "Technical Analysis: The Complete Resource for Financial Market Technicians", Financial Times Press, ISBN: 9780133093506
8. Ou, J. and Penman, S., 1989 "Financial statement analysis and the prediction of stock returns", Journal of Accounting and Economics, vol. 11, no. 4, pp. 295-329, https://doi.org/10.1016/0165-4101(89)90017-7
9. M. Noma, 2010. "Value investing and financial statement analysis", Hitotsubashi Journal of Commerce and Management, vol. 44, no. 1, pp. 29-46, https://doi.org/10.15057/18701
10. Nissim, D., 2022, "Big Data, accounting information, and valuation", Journal of Finance and Data Science, vol 8, pp. 69-85, https://doi.org/10.1016/j.jfds.2022.04.003
11. Tibshirani, R., 1996. "Regression Shrinkage and Selection via the Lasso", Journal of the Royal Statistical Society B, vol. 58, pp. 267-288, https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

12. Zou, H. and Hastie, T., 2005. “Regularization and Variable Selection via the Elastic Net”, Journal of Royal Society B, vol. 67(2), pp. 301-320, https://doi.org/10.1111/j.1467-9868.2005.00503.x
13. Masri, S.F and Bekey G.A, 1980. “A global optimization algorithm using adaptive random search”, Journal of Applied Mathematics and Computation, vol. 7 (4), pp. 353-375, https://doi.org/10.1016/0096-3003(80)90027-2
14. Ball, R. and Brown, P., 1968. “An empirical evaluation of accounting income numbers”, Journal of Accounting Research, vol. 6 (2), pp. 159-178, https://doi.org/10.2307/2490232
15. La Porta, R., 1996. “Expectations and cross-section of stock returns”, Journal of Finance, vol. 51 (5), pp. 1715-1742, https://doi.org/10.1111/j.1540-6261.1996.tb05223.x
16. Lee, W. and Jiang, C. and Indro, D., 2002. “Stock Market Volatility, Excess Returns, and the Role of Investor Sentiment.” Journal of Banking & Finance, vol. 26, pp. 2277–2299, https://doi.org/10.1016/S0378-4266(01)00202-3
17. Baker, M. and Wurgler J. and Yuan Y., 2012. “Global, local, and contagious investor sentiment”, Journal of Financial Economics, vol. 104(2), pp. 272-287, https://doi.org/10.1016/j.jfineco.2011.11.002
18. Koppel, C., 2021. “Does Social Media Sentiment Matter in the Pricing of U.S. Stocks?”, University of St. Gallen, https://dx.doi.org/10.2139/ssrn.3771788
19. Dash, S. and Maitra, D., 2022. “The COVID-19 pandemic uncertainty, investor sentiment, and global equity markets: Evidence from the time-frequency co-movements“, The North American Journal of Economics and Finance, vol. 62, 101712, https://doi.org/10.1016/j.najef.2022.101712