University of Tartu
Institute of Computer Science
Informatics Curriculum

Igor Atsberger

Possibilities of usage of machine learning in stock trading

Bachelor's Thesis (9 EAP)

Supervisor: Vambola Leping

Tartu 2020

Possibilities of usage of machine learning in stock trading

Abstract:

The rapid development of information technologies has made a favorable conditions for the development of new ways of earning money. This happened through accessibility of machine learning and stock trading to usual people. Combination of these two technical phenomenons could potentially make a person rich if they succeed in construction of neural network that would predict stock prices with a certain degree of precision. The goal of this research is to create such model using freely-accessible technologies.

Keywords:

Machine learning, stock trading

CERCS: P170

Masinõpe kasutamise võimalused börsil

Lühikokkuvõtte:

Kiire infotehnoloogia areng tekitas soodsaid tingimusi uute tuluallikate tekkimiseks. See juhtus tänu niisugustele tehnoloogiatele nagu masinõpe ja sellistele nähtustele nagu aktsiate börsi saadavus tavainimestele. Nende kahe fenomeni kombineerimine võib potentsiaalselt teha inimest rikkaks, kui temal õnnestub luua sellist tehisnärvivõrku, mis täpselt ennustaks börsi aktsiate hindu. Käesoleva töö eesmärk on sellise mudeli loomine vabalt kättesaadavate tehnoloogiate abil.

Võtmesõnad:

Masinõpe, börs

CERCS: P170

Contents

Introduction

The means of getting resources for the needs of people in our society is money. Usually people get money in exchange for their time. This is what is called work. If people were offered money for free, most of them would probably accept the offer.

In our society most basic things like that people need to live like food and house are acquired with money. Most people do hourly-paid work in other people's companies and those who do so usually work 40 hours a week, which makes up almost a quarter of total time each week. Many people would have spent this time some other way if they did not need the money that they received.

One of the ways that sometimes is believed to be a good way to make a fortune is playing on the finance market, which with the development of the internet and digitization also went digital and with that more accessible to people. There are lots of companies offering platforms for trading on the stock market or online courses that would teach how to be a successful trader. You just need to guess if the price of certain assets or currency is going up or down. Only two options, what could be that hard? Obviously free cheese can only be in a mousetrap. But many financial market experts state that if you dedicate enough money to study everything is possible.

Another field of technology that has seen a massive development recently is artificial intelligence. And in addition to that it also became more accessible to ordinary people because: open-source projects that allow usage of artificial intelligence such as Tensorflow and Keras along with ease of getting proper information and lessons on programming online and advancement of hardware that would be required for these programs using neural networks to run and people's accessibility to it. All that is needed to create neural networks and potentially make computers think instead of its owner is to own one, not even the most powerful.

So for many people who have heard about these two phenomenons a justified question occurs: why can neural networks be used on the financial market to make money?

The goal of this bachelor's thesis was to find a way to create a program that would be able to earn money off the stock market using machine learning or at least to find out if it is possible. In order to achieve that the documentation of machine learning framework was studied along with basics of stock trading.

The first chapter gives an overview of what the stock market and machine learning is. It names base principles those two technologies work by and related problems.
The second chapter describes methods used for development, such as choice of data, indicators and how models were programmed.
The third chapter describes the results and efficinity of models.
The fourth chapter summarizes the results of the work done and suggests future improvements.

1. Background

1.1 Stock trading

Generally speaking, stock trading is a long-known way of making money off predicting prices for valuable materials such as oil, gold, etc, currencies, companies' shares and so on. Although the details and rules of different kinds of stock trading methods differ, they are all based on prediction of what the financial instrument (stock or currency) price will be in one or another way. These predictions are made using technical or fundamental analysis. [1]

1.1.1 Fundamental analysis

Fundamental analysis "studies everything from the overall economy and industry conditions to the financial condition and management of companies". It requires analysis of data from different sources, like news, companies background, plans, often human-written text (e.g. news).[2] For the current level of development of machine learning techniques it is a hard task and humans would most likely outperform machines, but they would lose in terms of speed of data analysis and decisions making. [3] The creation of a machine learning model for this approach sure is possible, but it requires implementing a natural language analysis tool that would obtain information from for example news sites, and most likely a lot of human tuning and solid knowledge of economics. As a weak proof of the limitations stated above could be used the fact of absence of successful models when searched online. Yet again, the absence of proof does not prove contrary.

1.1.2 Technical analysis

Technical analysis is the method of prediction of prices of assets that is based on the implication that prices of the asset in future can be predicted from past price and volume graphs. Its only inputs are the stock price and volume, which allows more simple machine learning implementation for this method than in the case of fundamental analysis as those values are simple to extract, there is only few of them, they are usually accessible through the same trading platform or website where trading process is happening and are numeric unlike news articles from different resources.[2] Examples of technical analysis techniques are

moving average, support and resistance, trend lines, etc. There are also known patterns of price chart that are used to predict price movement.[4]

There are also indicators for technical analysis that in essence are results of application of some formula to graph. The indicators are meant to transform information in such a way that they help to predict price movements somehow. Each indicator was created to be used in a certain way, in other words there are usage instructions that go with each indicator. Usually the provider of a trading platform already includes indicators (results of calculations using this formula) into information that he provides.[5]

1.1.3 Problems

The whole topic of trading may seem controversial to most people - many call it gambling because of how unpredictable it is. [6] Most of those who try to make money off stock market fail, only small percent is successful (day trading is discussed here, not long-term investing), various sources without any information on where the statistics come from state that only around 5-15% of day traders are successful, others lose money. [7] The fact of someone's personal interest in statistics on how profitable trading is does not produce much confidence in those statistics. Sources that disclose any statistics have something to gain from their actions - trading platforms want to attract people to trade with them, professionals want to keep newcomers (potential rivals) off the market, and some people just lie about their profits to show off and make themselves look better than they are in reality. Lots of companies offer courses, sometimes even free or for a share of the student's profits (offering a kind of guarantee of quality of education), that are said to teach how to be a successful trader in several months or weeks. Opinion that trading requires a solid background in economics and very good knowledge of topic along with years of experience is also very common. Sadly, when money comes into play, the truth often bends in favor of financial interests of certain people.

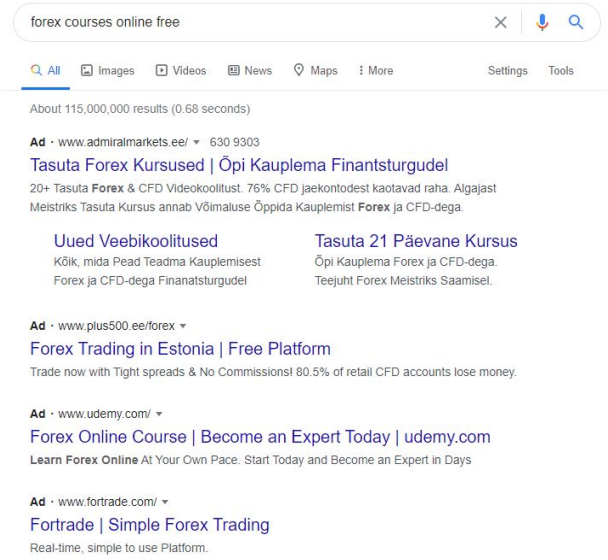

Fig 1. *Google results for search "forex courses online free". 115000000 results, first four are paid advertisements.*

1.2 Machine learning

Machine learning is a way of implementing artificial intelligence to create a mathematical model that predicts the "output" based on "input" after seeing several (usually from tens to thousands) pairs of "inputs" and "outputs" in the same form as the real data. Process of learning does not require human interaction or any changes in programm code, but it consists of change of mathematical model (called "fitting") in such a way that the correct "output" would be predicted on "input", so basically it tries to guess how to make a model such that "input" would produce correct "output". It works with some level of approximation - it can not only predict based on the exact same "input" that it had seen, but also if the "input" is a bit different and includes patterns of the previous input. The extraction of patterns is done according to the algorithms that this exact learning model is built on, different algorithms are suitable for different tasks (e.g. finding faces on pictures, predicting weather, etc.). The "transformation" functions that change the model in the process of learning (guessing) can also be different and suitable for different usage cases. Machine learning model is mathematical, it finds functions to transform one set of numeric values (input) to another (output), therefore data that it handles should also be in numerical form, so for example the picture is represented in pixels values.

Currently artificial intelligence is only able to be used in narrow applications for which it should be trained. For example a model can be created that would be able to tell if there is a dog or a cat in the picture if it has seen enough pictures with dogs and cats (with labels), but it can't answer what the meaning of life is.

1.3 Fusion

What is important for the applications of machine learning is the fact that data based on which the predictions are made can be represented in a form that the computer understands - a series of numerical data. Simpler models do technical analysis, which involves feeding past prices data into the model, delimiting it into "past" (the input) and "future" (the output).

In recent years machines started overtaking humans in this field and now around 80% share of the market is controlled by automated programs that require no human interaction with them and make on their own, as stated Yun Li in CNBC article.[8] How exactly this number was calculated is not explained. Another proof of the topic's popularity is the amount of research papers for this topic in EBSCO Discovery database - as for January 1st 2020 the number of results for the search "(machine learning or artificial intelligence or deep learning or neural network) AND (stock market OR trading)" is above 14000.

There's a lot of amateur implementations as well, none of which shows miraculous results.[9][10] But to think logically, if it was that easy, everyone around would be rich (or more likely the stock market would lose its current form and would change extremely in an unknown way). As one of the users of Stack Exchange suggests, "It is speculated that breakthroughs in the field of probability theory has happened several times, but never shared. This [could be] due to its practical application in gambling.".[11] This can be also said about application of AI in stock trading: why would anyone reveal a secret of becoming rich?

There are also different market tools, in addition to most common trading, where traders can place "long" and "short" offers, there are options, on which usage of machine learning and its performance can differ. For example if the tool that is used is Forex [12], then the decision when to buy assets, in what amount and also when to sell must be done, which requires constant calculation as the price updates every several seconds. In case of binary options the

situation is more simple - trader has to correctly guess if the price of certain asset will increase or decrease between current moment and a certain moment in the future (for example one hour later), no extra intermediate calculations needed, but there may be some additional rocks underwater as smaller profitability. [13]

One of the more technologically primitive method of automated trading with computers is just usage of indicators with added conditions, but it is not quite machine learning. There is also an option to feed indicators data into a neural network while training and a neural network would make sense of given data on its own or to use indicators as a helper method for the algorithm that programmer would teach on his own (to act in a certain way corresponding to what indicator shows).

There sure are attempts to perform machine learning using fundamental data, mostly analyzing news [14], but it's amateur level. It can't be denied that there may be some extremely advanced techniques and implementation, but they are not for public access as they are a source of extreme income. Also these models would not quite be machine learning as machine learning models would not likely be able to find relationships between for example news and stock prices movement as there are too many factors that influence the price and unlikely there are technologies already that are capable of a task of this level of difficulty. Programs that would be able to do fundamental analysis would use predefined relationships between news content and what would be its influence on stock price, so they would most likely require human design, therefore they cannot be smarter than people they were designed by. Program, though, can outperform people in terms of the amount of data it can analyze in a period of time. So it can be used for automation of the process of decision making, and also it can take more data into account at once without losing details out of sight, for example give a prediction based on hundreds of news articles. Yet again, news articles appear with a threshold and by the time the article appears, its value would be lost, as the market situation favorable for opening a deal would be already lost.[3]

2. Methodology

The goal is to create a program (machine learning model) that would be able to make profits off the stock market and in order to achieve that would predict price movement of different tools. Different approaches to neural network model creation are used, such as different input data, including data produced with trading indicators.

As the data that predicted is numeric, regression machine learning methods will be used. There was an option to use a real brokerage platform API[1], for example Interactive Brokers [15]. It also has a demo account, so no real money is needed to test out the performance. But they have limitations that will slow down development drastically, such as the limited amount of data received and integrating would also need a lot of additional effort even though it is quite possible. It is the best way to show performance and the most real one, closest to real world use.

The goal of this research is to simulate real world conditions - data is received gradually over time, and with the coming of new data decision should be made - to open the trade or not to and in which direction. Model would simulate day trading as it is more suitable for pure technical analysis. The longer the period of investments the higher is influence of macroeconomics that technical analysis cannot handle.

2.1 Data

Getting data required - consistent and with high frequency is a difficult task - most of the providers found want hundreds of dollars for their services. [16][17] Data was retrieved using Tickstory software. [18] The software allows downloading of stock market and forex data in various formats including CSV[2] and it has a free version with all the functionality needed for this research. Data used for this research is every minute forex data for EURUSD pair for the last 5 years. It includes date and timestamp, open, high, low and close prices and volume.

[1]Application programming interface
[2]Comma separated value - a format used by Microsoft Excel and Pandas Module in Python

Some data is missing: if the data has a datapoint for every minute in the past five years, then there should be about 2628000 datapoints, but the given data has only 1867813, so about 30% of data is missing. It must be somehow handled for resulting models to perform better on the real world use. One option is to fill missing data points with the average of two nearest available points. Another option is to ignore the gap and act as if it doesn't exist and just consider two nearest available data points as the ones that are immediately next to each other in reality. The third option is to just use periods of time with no gaps (or with minimal gaps). There are its own problems in first two options: the first one makes the model "think" prices change over time more smoothly and with less noise than in reality, whereas the second one makes it act as if prices made bigger jumps. Analysis of data shows that the third option would drastically decrease the number of samples to learn and test on, therefore performance of models could suffer.

Data needs to be normalized, or else firstly - transformation functions will not work properly as they are designed to work with values of range from zero to one, [19] and secondly - data will be interpreted the wrong way - in stock trading the difference between the point of order opening and order closing is the key, not the starting and finishing price on their own.

The formula of normalization of each for the models that make predictions based on price is the following: *for each in sequence: each/sequence[0]-1*. In the resulting sequences absolute most of values, about 95% is under 0,001 and the rest are slightly higher than 0.01. Such low variability also has a negative effect on the performance of activation functions on which training of the model relies. For better performance each of the values in each sequence is multiplied by 100, so the resulting normalization formula is the following: *for each in sequence: (each/sequence[0]-1)*100*. This formula, though, results in values being roughly between -1 and 1. As it was said before, activation functions work best with data in range between 0 and 1. If this assumption is correct, formula *for each in sequence: (each/sequence[0]-1)*50+0.5* should be applied. The downside of this option is that for neural network prediction of price of 0.49 and 0.51 would seem close, where in reality it means that the direction of price movement is predicted mistakenly.

Also trading indicators would be used for some models. Indicator, again, is just a mathematical function applied to market data (price and volume).

2.2 Models

For prediction, the neural network model is created. Modern technologies, such as framework Keras[20] for Python programming language makes creation and usage of neural networks accessible to everyone and understandable even for people who are not scientists with years of experience and academy training. Yet it is clear that scientists will have way better results, keras just lowers the entry bar. This library allows creation of neural networks by simple invocation of pre-defined models, layers, activation functions and so on, so for simple (but well-working) neural networks the program code could be under 100 rows long. Compared to how complicated implementation of neural networks could be - defining mathematical functions it is a huge leap in accessibility. Keras actually is just an API which uses for example TensorFlow[21] as a backend.

Different neural network models with different layer types are tried. A model is a set of layers. Different models in this work use different layers, normalization functions, data inputs, number of epochs. The only resource whose consumption rises with increasing number of epochs is time taken for learning. It is a "cheap" resource in a sense that it does not require some special hardware and can be exchanged for better results. For each model such a number of epochs is chosen that allows the model to reach the state where there are no visible improvements in loss metric in several last epochs. The number of epochs is determined during trial and error.

2.2.1 Pure price based

The first model uses only price data from the past 24 hours (each minute) to predict the price one hour later after the last known (given to model) data point. For example it takes prices each minute from the noon of 17th of April up to the noon of 18th of April and it should predict what price the asset will have at 1 A.M. on 18th of April. If the data is used as described before, the number of training samples (24 hours + one point one hour later) is around 1000 and number of test samples is around 250. There is also an option to make samples overlap (sometimes called "the window method") that would not be used because of

hardware limitations on certain computer, more precisely 16GB of RAM is not enough for the implementation, which was discovered with trial and error method. This method is close to how binary options work - trader needs to predict how price is going to change compared to the current over a certain period of time, for example one hour.[13]

The model uses four lstm layers with numbers of neurons 25, 50, 50, 25 respectively and one dense layer. Lstm is often used for predicting time series.[22][23][24] These numbers were again received in experimental way, limited by the hardware.

Relu is used as an activation function as it was the most effective activation function from personal experience. Initial number of epochs is randomly set to 20, but is planned to increase if the learning curve does not flatten in 20 epochs.

2.2.2 Pure price based variant two

This is the same as the previous model, but each layer has 5, 10, 10, 5 neurons respectively and it uses 100 epochs. It is made because of the assumption that shifting usage of resources from number of neurons to number of epochs could improve performance. Also the big number of neurons could overcomplicate the model, which could result in overfitting.

2.2.3 Price + volume based

This model is similar to the first one, but it also uses trading volume as an input which in theory should be helping in price prediction as most of the trading indicators use volume in its formulas and it is also considered an important metric when making a trading decision [25]. In order to achieve that a principle that is called multiple regression is used. [26] [27] It means that instead of sequence of prices, the input consists of sequence of pairs price+volume and the output is single price.

Taking into consideration the fact that there are more features, it is possible that the model would take more time to find relationships between input and output, therefore, the number of epochs is raised to 50. As ‘relu’ normalization function can not be used with multiple regression, default activation function is used instead. This was found the practical way as using relu resulted in error.

2.2.4 Bollinger Bands

Bollinger Bands named after its creator is a popular indicator in trading community. [28] This indicator consists of three parts - moving average and two bands that respectively are moving average plus and minus two standard deviations. Moving average is used only for calculation of bands. Recommended usage is to open sell position when price is above the upper band and open buy position when price is below the lower band. The period the moving average is calculated on can vary. The shorter the period, the more sensitive the indicator is and the period for which the orders are put are shorter as indicator predicts growth of fall for a shorter periods of time.[29][30] For this particular case the period of 20 data points, meaning minutes is chosen.

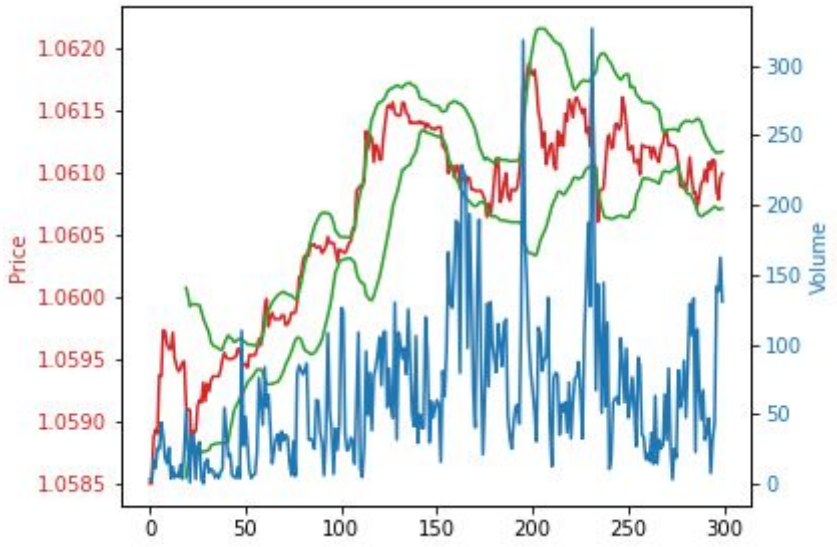

Fig 2. *Example of plot depicting Bollinger Bands (green). Plot produced with matplotlib and using data used for creation of models by the author.*

2.2.5 MACD

MACD (*Moving Average Convergence Divergence*) is another popular trading indicator. [28] The indicator consists of two lines. The first one, called the MACD line is 26 period exponential moving average subtracted from 12 period moving average. The second line is a 9 period exponential moving average of the MACD line. MACD is meant to be used this way: when MACD crosses the signal line from below it signals that the price is going to grow for some time and it may be a good time to buy. When the signal line is crossed by the MACD line from above it is a signal to sell. The bigger the angle between lines around the

intersection point, the stronger the signal is.[31] Again, as with Bollinger Bands, periods are important as relative measures, in other words periods can be increased or decreased, but with the same multiplier. The rule of sensitivity described in Bollinger Bands also applies here.
As periods in the current case are minutes, the periods are increased ten times each in order for indicator not to be too sensitive. Both MACD and signal lines are in unsuitable range - from about -0.005 to 0.005 if calculated on unnormalized price data. So the data sequences are multiplied by 100 and 0.5 is added to each element of resulting sequences. As a result, the values fit into range from 0 to 1.

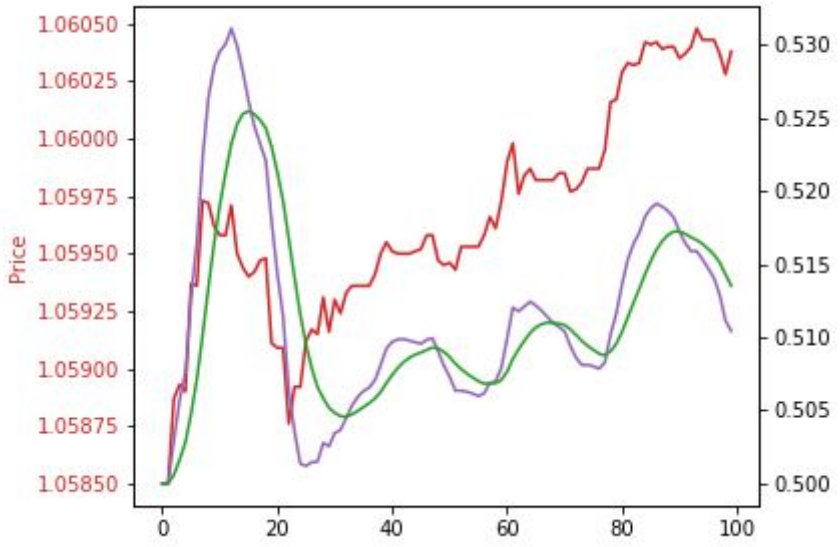

Fig 3. *Plot of MACD indicators. The MACD line is purple, the signal line is green. In this particular case most of the signals "tell the truth".*

2.3 Evaluation criteria

As was repeatedly suggested in Tartu University course "Machine Learning" the classical 80/20 data split is used - 80% of data is training data for models and 20% of data is for testing.

For evaluation the test results (neural network predictions) are compared to the actual values from the data. For each prediction the formula *(prediction - last_data_point)/(target -last_data_point)* is used. *Target* here is the value that should have been predicted (ideally *target* should be equal to *prediction*). The sign that this formula result has shows if the direction of price movement was predicted correctly. Then the percentage of correct direction

guesses is calculated. The threshold is 50%. Getting 50% of guesses correctly is the same result as for example tossing a coin to guess where the price is going to move as there is only two options to choose from and if infinite tries to randomly guess is made, the accuracy tends to zero.

Additionally, built-in metrics from keras are used to analyze the learning curve through epochs and evaluation of performance on test data. The same metric - Mean Square Error loss metric is used in all the models so that comparison of models' performance would be more obvious. [32]

3. Results

The normalization formula *for each in sequence: (each/sequence[0]-1)*50+0.5* produced slightly better results if test evaluation loss is compared than *for each in sequence: (each/sequence[0]-1)*100* for the first model, therefore it will be used for all the later models.

3.1 Pure price based

Results of several identical tests on first model showed price movement direction precision of 51, 46, 49 and 56 percent, which is close to precision of random picking out of two options. Although more tries needed to calculate confidence interval, it seems that this model cannot be of any use and for example putting a monkey in front of two bananas labeled “up” and “down” respectively and letting it choose one would be equal to usage of the given model.

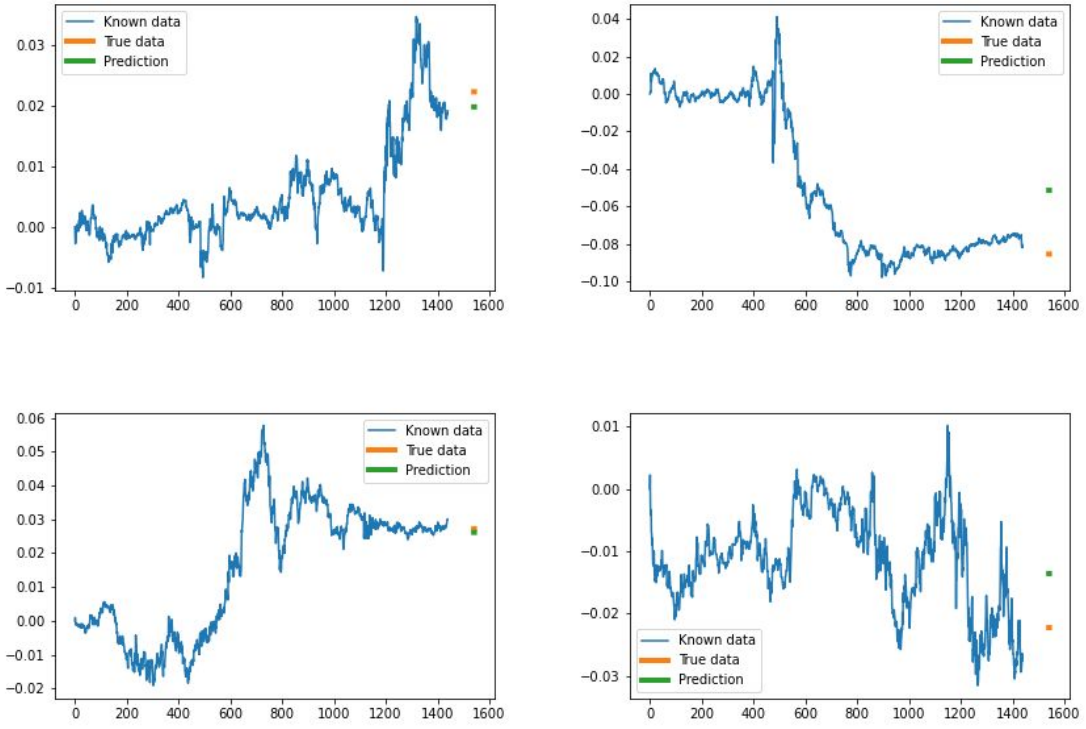

Fig 4. *Several pseudo-randomly chosen plots of test results of the first model*

At least it is clear that the error that the model made in its prediction is in adequate range, e.g. there is no prediction that differs from real change 100 or 1000 times, therefore the correctness of inputs can be implied. The loss during testing is 0.006.

3.2 Pure price based variant two

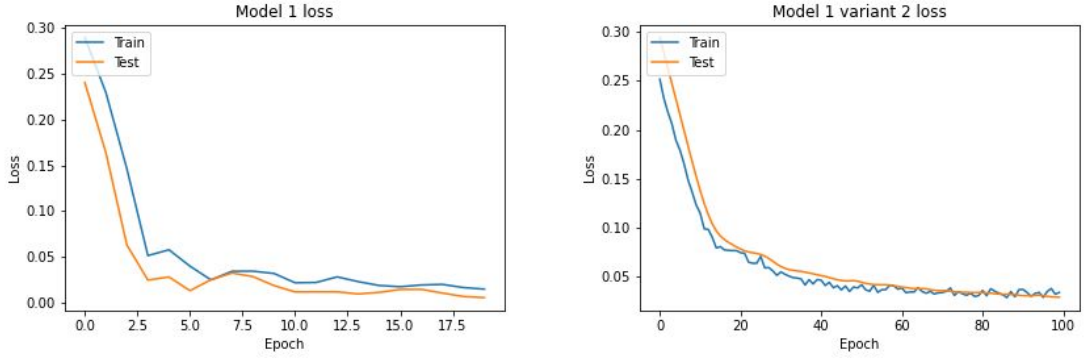

Fig 5. *Plots of loss metric change with number of epochs for Model 1 and Model 1 Variant 2*

The second variant of this model has comparable performance, about 47%-53% each time the testing is about 0.01, which is worse than the original Model 1. It could be stated that in this particular situation reducing the number of neurons and increasing the number of epochs did not improve the performance and vice versa made it worse.

It is visible that 20 epochs was enough for model 1 to stop improving, so this number of epochs is enough.

3.3 Price + volume based

Adding volume to data gave no positive changes in accuracy of predictions. Model still guesses the direction of movement of price with accuracy around 50%. Loss for test evaluation is 0.008.

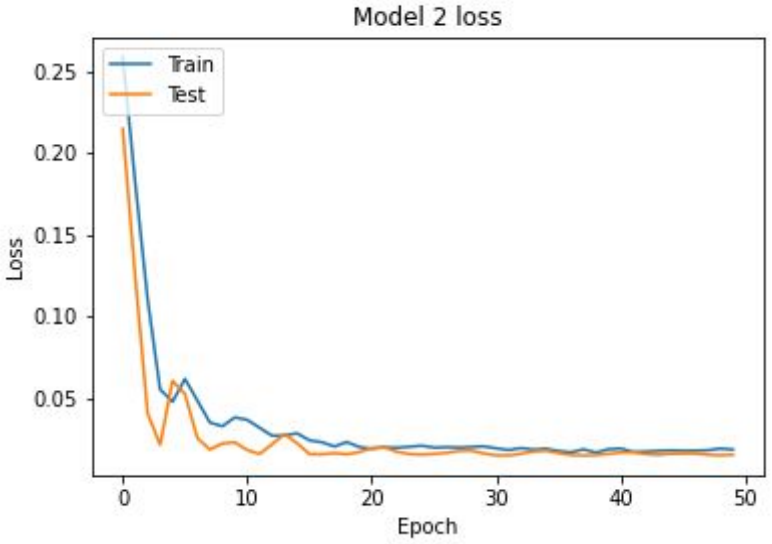

Fig 6. *Plots of loss metric change with number of epochs for Model 2*

3.4 Bollinger Bands

Bollinger bands performed better than any aforementioned models with test loss about 0.004, but there was no breakthrough about the accuracy of predictions of price movement - also about 53-54 percent.

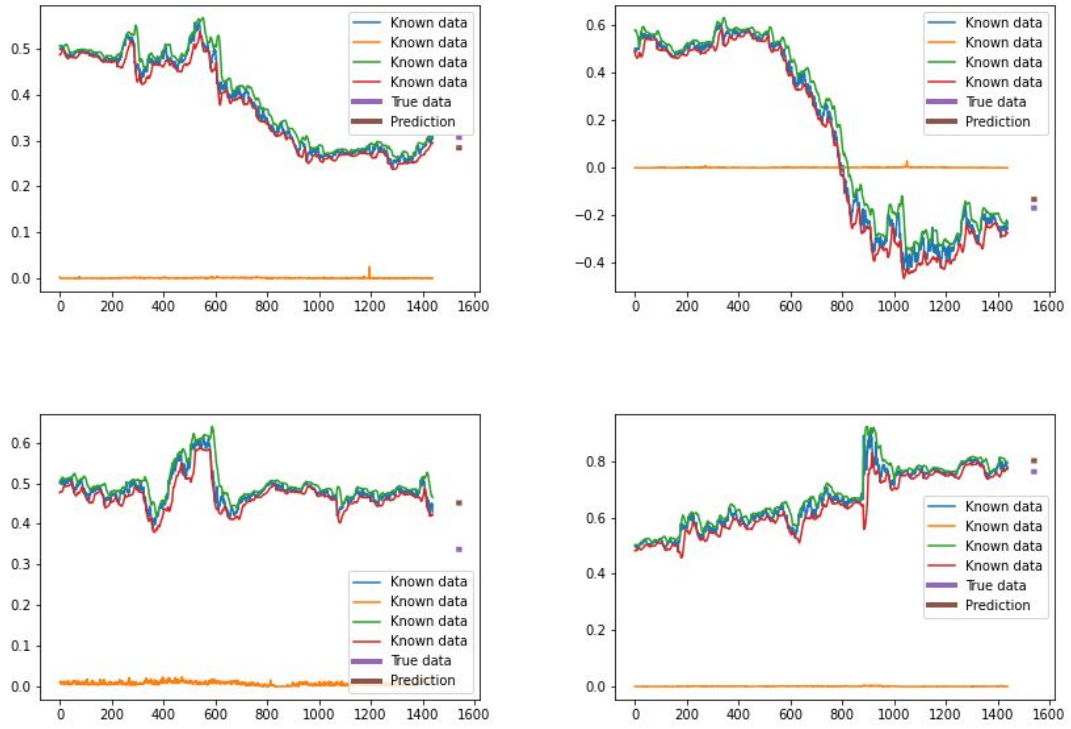

Fig 7. *Several pseudo-randomly chosen plots of test results of the Bollinger Bands Model*

From the plots on figure 7 the tendency could be noticed that the model tries not to predict value too far away from the last price data point in comparison to what real price is. Mean absolute difference of real target price compared to buy point is 0.035, predicted price has on average 0.017 absolute difference from the last price data point that was fed to the neural network. Similar situation is with the standard deviation - for a list of absolute differences from buy point price to real price this value is 0.044, whereas for a list of absolute differences from buy point price to predicted price is 0.036. Perhaps it is a way for a model to get higher accuracy as quite often the target of prediction would indeed not differ much from the last known price as seen in plot examples and making extreme assumptions would mean bigger errors.

Another assumption is that the greater the difference between last known price and prediction is, the more confident the prediction is and probably adding some threshold would improve prediction quality.

Threshold	Precision %	Number of samples
none	53.82	249
0.005	50.54	184
0.01	53.08	130
0.015	56.52	92
0.02	54.39	57
0.025	53.49	43
0.03	46.88	32
0.035	48.0	25
0.04	60.0	20
0.045	61.54	13
0.05	71.43	7
0.055	100.0	5
0.06	100.0	5
0.065	100.0	5
0.07	100.0	4

Fig 8. *Precision of Bollinger Bands model with different thresholds on differences between last known price and predicted price*

In this particular case using high threshold improved test results, repeating the same process again - making model learn from zero and then predict - showed that these great results were just a statistical error. Shown on figure 10.

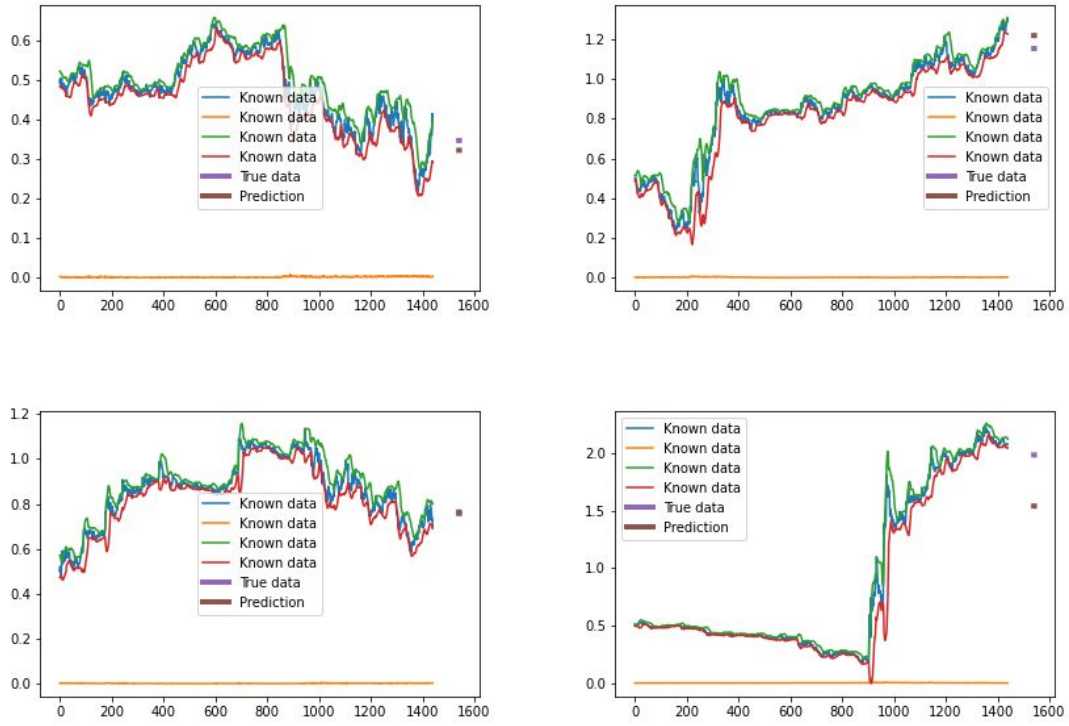

Fig 9. *Plots of test cases from the set of filtered cases with the highest threshold.*

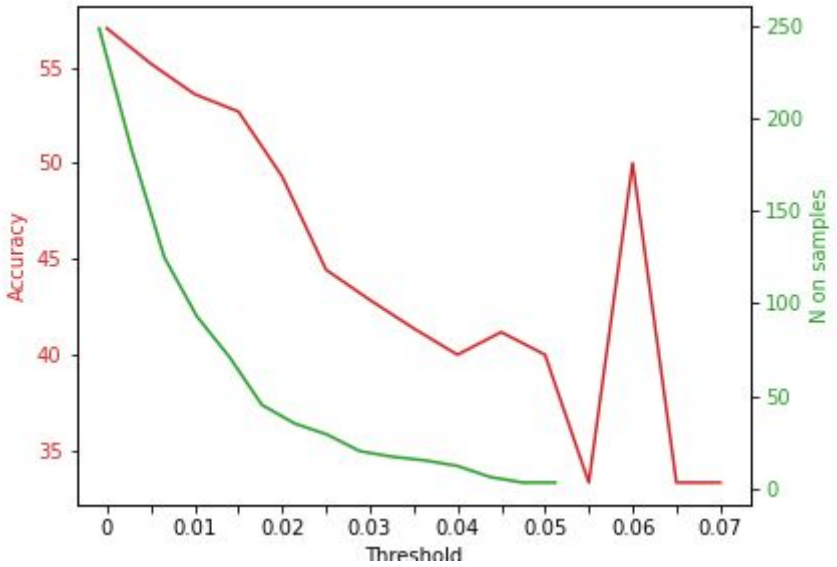

Fig 10. *Precision of Bollinger Bands model with different thresholds on differences between last known price and predicted price - second attempt.*

3.5 MACD

MACD did not stand out too much - during testing the loss is 0.0045, comparable to the performance of Bollinger Bands. Tendencies similar to the ones noticed in Bollinger Bands case can be noticed here - mean and standard deviation for the absolute of predicted price minus last known price is smaller that the same values for real target price, 0.023 and 0.039 against 0.034 and 0.042 respectively. Another trend is an increase of accuracy when raising the prediction difference from the last known price threshold. The plot on figure 11 represents it.

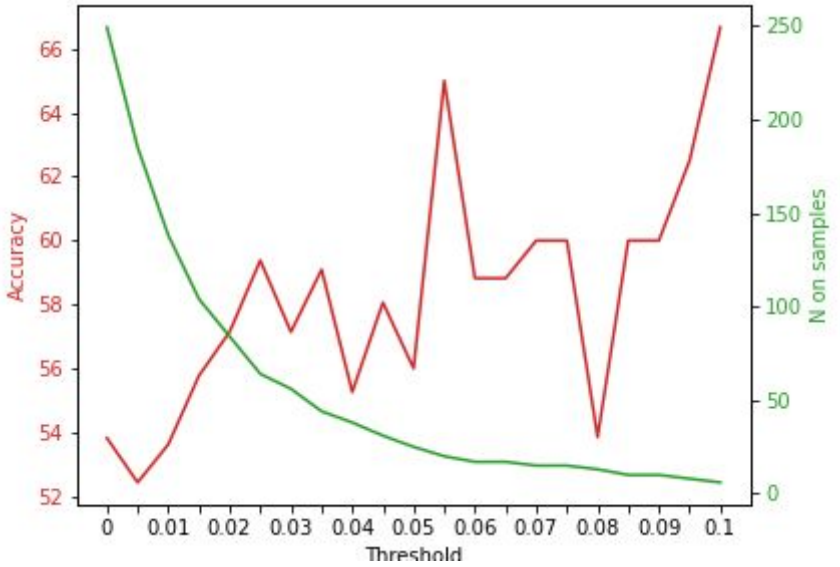

Fig 11. *Precision of MACD model with different thresholds on differences between last known price and predicted price*

The increase is not steady, similarly to the Bollinger Bands case, therefore success using this threshold is not guaranteed.

4. Summary

Performance of the models is rather weak, the result is generally just slightly above the random threshold of 50% - about 53-55%. Although much more testing is required, for example multiple repeated running of the model in order to produce confidence interval.

There is a chance that existing models' performance could be improved with the application of more trial and error in input data, e.g. using different combinations of indicators, different normalization, choice of time windows and data density etc. Creating something useful in this topic requires much more specific skills and knowledge in machine learning and artificial intelligence, economy and stock trading. Amount of work needed is also far beyond of the amount that is usually planned for bachelor's thesis.

The created models can be tested to work as assistive tool for the cases where trader sees an entry point to buy or sell assets. The way that these models were tested in current work differ from the likely behaviour of real human trader - the model was forced to give its prediction in all the market conditions, but in real trading trader most of the time would refuse to make any deal. There is a chance that the model would be able to help potential trader make any profit if he uses it to predict future only when he by himself sees a potential trading opportunity. Such tests were not done due to lack of trading skills.

Different loss functions could be used in training of given models. Current loss which is Mean Squared Error does not distinct between the cases where the prediction is off by the same amount of points, but in one case the direction is guessed correctly and in the other it is not, therefore first case would mean losses to a person whose money the model uses for trading and the other would mean profit. The possible different model could mean bigger loss of score if the direction was guessed incorrectly than when it was guessed correctly and therefore to minimize error the model would try harder to avoid incorrect guesses of direction.

Even though the profitability of created models is highly doubtful, the program can be created to integrate into trading platform API. Such a program would require some kind of decision making mechanic along with confidence evaluation - when to open a trade and when to close it. For example the program could start considering if it makes sense to open a trade based on a script that uses traditional trading indicators with human-written decision making process, for example using a MACD indicator when two its lines cross.[3] Or it could be done vice versa - each minute the prediction using a neural network model is made and if the scripted indicators approve, the trade is made. This would again require more skills in machine learning and trading. The simpler to implement method, which would not require that much extra programming could be to just open trades only when the predicted change of price is higher than some threshold.

Money management techniques can also be implemented with python if a real trading account is used. Such data would be retrieved from service provider through API.

All in all, the results if proven to be correct during additional testing exceeded the expectations that were rather skeptical. Even though only by 3 or 4 percent.

Sources

[1] Corporate Finance Institute. What is the Stock Market and How it Works https://corporatefinanceinstitute.com/resources/knowledge/trading-investing/stock-market/ (26 January 2020)

[2] Majaski C. Fundamental vs. Technical Analysis: What's the Difference? *Investopedia,* 14 April 2019 https://www.investopedia.com/ask/answers/difference-between-fundamental-and-technical-analysis/ (26 January 2020)

[3] Stewart M. The Limitations of Machine Learning *Towards Data Science,* 29 July 2019 https://towardsdatascience.com/the-limitations-of-machine-learning-a00e0c3040c6 (8 May 2020)

[4] Hayes A. Introduction to Technical Analysis Price Patterns *Investopedia,* 25 June 2019 https://www.investopedia.com/ask/answers/difference-between-fundamental-and-technical-analysis/ (26 January 2020)

[5] AnyChart.Com. Technical Indicators Mathematical Description https://docs.anychart.com/Stock_Charts/Technical_Indicators/Mathematical_Description (22 February 2020)

[6] Yell T. The Similarities Between Day Trading and Gambling *The Balance*, 3 October 2019 https://www.thebalance.com/the-striking-similarities-between-trading-and-gambling-1345200 (8 May 2020)

[7] Rolf Scientist Discovered Why Most Traders Lose Money – 24 Surprising Statistics *Tradeciety* https://www.tradeciety.com/24-statistics-why-most-traders-lose-money/ (19 April 2020)

[8] Li Y. 80% of the stock market is now on autopilot. *CNBC,* 29 June 2019 https://www.cnbc.com/2019/06/28/80percent-of-the-stock-market-is-now-on-autopilot.html (26 January 2020)

[9] Lu B. Machine Learning for Stock Market Investing. *Medium,* 31 January 2019 https://medium.com/datadriveninvestor/machine-learning-for-stock-market-investing-f90ad3478b64 (26 January 2020)

[10] Xu J. How To Use Machine Learning To Possibly Become A Millionaire: Predicting The Stock Market? *Towards Data Science,* 30 August 2019 https://towardsdatascience.com/how-to-use-machine-learning-to-possibly-become-a-millionaire-predicting-the-stock-market-33861916e9c5 (26 January 2020)

[11] StackExchange. How can I go about applying machine learning algorithms to stock markets? *StackExchange* https://quant.stackexchange.com/questions/111/how-can-i-go-about-applying-machine-learning-algorithms-to-stock-markets (18 April 2020)

[12] Chen J. Forex Trading: A Beginner's Guide *Investopedia,* 16 March 2019 https://www.investopedia.com/articles/forex/11/why-trade-forex.asp (8 May 2020)

[13] Smith T. Binary Option *Investopedia,* 16 March 2019 https://www.investopedia.com/terms/b/binary-option.asp (8 May 2020)

[14] Braun M. This Machine Turns Trump Tweets into Planned Parenthood Donations *Medium,* 6 February 2017 https://medium.com/@maxbraun/this-machine-turns-trump-tweets-into-planned-parenthood-donations-4ece8301e722 (8 May 2020)

[15] IB API | Interactive Brokers U.K. Limited https://www.interactivebrokers.co.uk/en/index.php?f=40022 (8 May 2020)

[16] Real-time, tick by tick data for Stock, Futures, and Commoditie https://unibit.ai/pricing (19 April 2020)

[17] AlgoSeek - Historical Institutional Intraday US Market Data https://www.algoseek.com/ (19 April 2020)

[18] Tickstory - Free Historical Tick Data & Trading Resources https://tickstory.com/ (19 April 2020)

[19] Jaitley U. Why Data Normalization is necessary for Machine Learning models *Medium,* 8 October 2018
https://medium.com/@urvashilluniya/why-data-normalization-is-necessary-for-machine-learning-models-681b65a05029 (8 May 2020)

[20] Keras: the Python deep learning API
https://keras.io/ (8 May 2020)

[21] TensorFlow - An end-to-end open source machine learning platform
https://www.tensorflow.org/ (8 May 2020)

[22] Mwiti D. Using a Keras Long Short-Term Memory (LSTM) Model to Predict Stock Prices *KDnuggets*, November 2018
https://www.kdnuggets.com/2018/11/keras-long-short-term-memory-lstm-model-predict-stock-prices.html (8 May 2020)

[23] Brownlee J. Time Series Prediction with LSTM Recurrent Neural Networks in Python with Keras *Machine Learning Mastery*, 21 July 2016
https://machinelearningmastery.com/time-series-prediction-lstm-recurrent-neural-networks-python-keras/ (8 May 2020)

[24] Brownlee J. How to Develop LSTM Models for Time Series Forecasting *Machine Learning Mastery*, 14 November 2018
https://machinelearningmastery.com/how-to-develop-lstm-models-for-time-series-forecasting/ (8 May 2020)

[25] Bourquin T. Using Volume To Predict Price Movement *Money Show*, 29 November 2007
https://www.moneyshow.com/articles/daytraders-4676/ (8 May 2020)

[26] Del Valle Vega R. and G Rai A. Multivariate Regression | Brilliant Math & Science Wiki *Brilliant*

https://brilliant.org/wiki/multivariate-regression/#multiple-regression (8 May 2020)

[27] Brownlee J. Multivariate Time Series Forecasting with LSTMs in Keras *Machine Learning Mastery*, 14 August 2017 https://machinelearningmastery.com/multivariate-time-series-forecasting-lstms-keras/ (8 May 2020)

[28] eToro. eToro's guide to technical analysis tools & trader's lingo *eToro*, 20 February 2020 https://www.etoro.com/blog/trading-essentials/a-guide-to-popular-technical-analysis-tools/?gclid=CjwKCAjw7e_0BRB7EiwAlH-goMic9NSHdaCBpvfl3bVQPqncnFknHZ-WKQDEEpECGgGDvTyI8m7Q3BoCQocQAvD_BwE&utm_medium=SEM&utm_source=70285&utm_content=0&utm_serial=ROE_NB_DSA_EN_70285|AG_71983897393|KW_|MT_b&utm_campaign=ROE_NB_DSA_EN_70285|AG_71983897393|KW_|MT_b&utm_term=&gclid=CjwKCAjw7e_0BRB7EiwAlH-goMic9NSHdaCBpvfl3bVQPqncnFknHZ-WKQDEEpECGgGDvTyI8m7Q3BoCQocQAvD_BwE (8 May 2020)

[29] Bollinger J. John Bollinger's Official Bollinger Band Website https://www.bollingerbands.com/ (8 May 2020)

[30] Brenyah B. Setting up a Bollinger Band with Python *Medium*, 13 January 2018 https://medium.com/python-data/setting-up-a-bollinger-band-with-python-28941e2fa300 (8 May 2020)

[31] Posey L. Implementing MACD in Python *Towards Data Science*, 30 March 2019 https://towardsdatascience.com/implementing-macd-in-python-cc9b2280126a (8 May 2020)

[32] Keras API reference -Regression losses https://keras.io/api/losses/regression_losses/#meansquarederror-class (8 May 2020)

Appendix

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, **Igor Atsberger**,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose

   Possibilities of usage of machine learning in stock trading,

   mille juhendaja on **Vambola Leping,**

reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 3.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.

3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.

4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Igor Atsberger

08.05.2020