Mateusz Dadej
III year SS1 Economics Theory of Enterprises

APPLICATION OF ENSEMBLE GRADIENT BOOSTING DECISION TREES TO FORECAST STOCK PRICE ON WSE

Key words: equity investments, artificial intelligence, machine learning, algorithmic trading strategy, gradient boosting

Introduction

For a long time, academic theorists considered financial markets as informationally efficient, in the sense that current prices reflect every available information about the given security. Yet, since the crash of 1987, economists beliefs about the efficient market hypothesis has begun to change. Along with extensive empirical evidence and ingenious theoretical publications, the nonexistence of EMH gained in favor amid academics. Economic work, considered one of the most reasonable assertion against EMH, done by J. E. Stiglitz and Sanford J. Grossman[1], states that markets cannot be informationally efficient, for there is a cost associated with obtaining information, hence, the informed investors would receive no benefit. Due to mentioned, Grossman-Stiglitz Paradox, and other evidence, EMH cease to be prevailing notion on market efficiency. Above regime shift in academia, gave a theoretical permission to seek for methods to achieve an alpha in a long term. It was also, partly, a motivation for author to undertake following research.

The stock market is a very complex system and as well as other financial time series, is considered nonlinear and chaotic. Albeit, as emphasized before, to some extent predictable, despite of not so solid scientific consensus in this regard. As a result of its complexity, caused by countless factors influencing it, many scientists and practitioners of investment industry tend to use quantitative methods to analyze financial markets or support their decisions. Additional factor, that increased significance of quantitative finance in recent years, was emerging of machine learning. In the context of forecasting time series in finance, machine learning offer substantial enhancement in respect to more conventional methods, as a result of its lack of linear relationship assumption between

[1] J. E. Stiglitz, Sanford J. Grossman , On the Impossibility of Informationally Efficient Markets, „ The American Economic Review” 1980, Vol. 70, No. 3.

variables. Furthermore, machine learning algorithms are appreciated for its ability to identify hidden patterns and relations in analyzed data, thus demonstrating high accuracy of prediction. One of the algorithms that author want to highlight in this paper is ensemble of classification and regression trees (CART), which will be boosted with the use of extreme gradient boosting framework, described by T. Chen and C. Guestrin[2]. For the purpose of this article, the algorithm will be thereinafter referred to by the author as the XGBoost model, which is also commonly called across scientific literature.

The main objective of this research is to utilize machine learning model based on ensemble of boosted decision trees algorithm to predict direction of share prices. The model will be applied to polish stock market and later evaluated with standard methods and a Monte Carlo simulation.

1. Research data and methods

As stated before, presented algorithm belongs to the field known as machine learning. It is a study of how particular mathematical models gather data and analyze it in order to execute prespecified task. Machine learning is generally divided into two branches, supervised learning and unsupervised learning. During first one, an algorithm receive input data and labels in the form of correct output data, which enable it to choose parameters, that can generalize the unobserved data as good as possible. In opposite to supervised, during unsupervised learning the analyzed data lacks the correct output. Algorithm tries to discover similarities and identify new correct output.

The most prevailing process in predictive modeling, which will be also conducted later, consists of four subsequent parts. Transforming and cleansing gathered data, choosing significant variables or if necessary, feature engineering, building predictive model and at the end, model evaluation. Every phase above might be carried out with use of many different technics.

The primary data analyzed in this research is Bank Handlowy S.A. stock closing price and its volume on Warsaw Stock Exchange from IPO to 30 November 2018. The main rationale behind selection of a mentioned company was a long history of its stock quotations, which is important since machine learning algorithms are the more effective the more data they will learn. Another reason is a relatively weak long term uptrend of stock price. Since the IPO in 1992, stock price is now around 5 times higher and only 49,7% of daily stock returns were positive. Therefore, there will not be any significant uptrend bias in predictive model. The stock data is not adjusted for

[2] Guestrin C., Chen T., *XGBoost: A Scalable Tree Boosting System*, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, 2016.

dividends paid, right issues and share splits, however this will also not influence the outcome of our analysis .The dataset has been gathered from www.bossa.pl.

With the use of raw data described above, it is possible to derive following features included in final dataset. First feature obtained, which is return from next 5 trading days is used for making response variable with the use of one hot encoding. Numerical values were converted into categorical ones, such that positive returns are encoded as 1s and negatives as 0s. Given the character of the feature above, it is clear now that the model will perform classification task, predicting if the stock price will go up (1) or down (0). Additionally, following feature engineering has been done to extract explanatory variables. Closing prices and volume has been lagged in order to create 3 variables of their values from t-1, t-2, t-3 to include momentum of the price in model. 14 commonly used technical indicators were derived from closing prices data, these are: 15, 30 and 60 days moving average, 30 days exponential moving average, Bollinger bands, moving average convergence divergence, relative strength index, slow and fast stochastic oscillator and stochastic momentum index.

In herein research, author will apply machine learning algorithm based upon classification and regression trees. CART basic concept is binary recursive partitioning procedure, which builds a model in a decision tree scheme. The division criterion is to group observations with similar response variable. As the name suggests, CART model is able to perform a regression and classification task, although, later on it will be used only for classification purposes. In order to increase accuracy of the model, a gradient boosting technique was used in this research, which combines a model from an ensemble of many weak models. Extreme gradient boosting is one of the most efficient implementation of gradient boosting. To build a extreme gradient boosting regressor the trees that minimize a loss function are chosen. A loss function is composed of two factors: an error rate that is calculated over a validation data set and a regularization factor to avoid overftting the model[3].

2. Empirical application

Before utilizing forementioned model, it is crucial to randomly divide dataset into two parts, testing and training data. This will allow to better evaluate model and consider it able to predict out-of-sample data, which is fundamental objective for a predictive model. Standard division is 70% of training and 30% of testing data. Given that, the model will learn based on

[3] B. Stearns, et.al, Scholar Performance Prediction using Boosted Regression Trees Techniques, „ESANN 2017 proceedings” 2017.

4401 observations and will be further evaluated with 1886 observations each with 34 explanatory variables.

With the final dataset, a model can be learned based on input data. After a series of attempts, a proper model tuning was also performed with discretionary specified final parameters that are generally divided into three main parts. General parameters, describing which underlying model is being used for boosting. As mentioned before, it is a tree based model. Booster parameters, specifying parameters of boosting, such as learning rate, depth of decision tree, step size shrinkage, etc. And the last one, learning task parameters, specify what task should model fulfill, regression or classification and on which evaluation metric should model depend during learning phase.

Final model is hard to visualize, although it can be analyzed to some extent on 3 metrics describing importance of particular features in model. First one is gain, which is average contribution, based on the total gain of this feature's splits across decision trees in percents. Cover metric is a share of observations that was described by given feature during learning phase. Frequency is the percentage representing the relative number of times a particular feature occurs in the trees of the model. Gain is often considered the most relevant metric. Table below represents values of a metrics for the 8 most important features.

Feature	Gain	Cover	Frequency
60 days SMA	0.0794874	0.0615013	0.0584353
MACD	0.0769266	0.0569931	0.0665128
30 days SMA	0.0569233	0.0499567	0.0465267
MACD’s signal line	0.0534922	0.0601962	0.0627279
Volume of the previous day	0.0510576	0.0507356	0.0553427
volume of the previous third day	0.0501035	0.0369588	0.0559889
RSI	0.0495859	0.0467054	0.0501269
Closing price of current trading day	0.0495809	0.0450946	0.0437111

Table 1 various metrics of model features importance. Own elaboration.

During the process of developing the code and applying quantitative methods contained in herein research, the author have made extensive use of R programming language and its libraries, especially xgboost, dplyr and TTR.

3. Model valuation

Because predictive model main objective is to predict out of sample data, it is essential to evaluate its predictive ability on such dataset. That is why we divided main dataset earlier. Testing dataset consists of 1886 observations each with 22 explanatory variables, without response variable for it will not be available while applying model in practice.

Table number 2 represents confusion matrix of prediction results. It is a two-dimensional matrix, indexed in one dimension by the true class of an object and in the other by the class that the classifier assigns[4]. Additionally, three ratios were calculated on the basis of confusion matrix. True positive rate, also called sensitivity, is a sum of false negatives divided by sum of true positives and false negatives. It measures the probability of detection, in this case probability of correct buy signal. True negative rate, measures proportions of correctly identified negatives, which can be also understood as a probability of correct sell signal. Accuracy, the most synthetic one, is proportion of correctly identified outcomes.

Total population = 1886		Actual class	
		1	0
Assigned class	1	True positive = 703	False positive = 280
	0	False negative=233	True negative = 670
Accuracy = 0.7279958		True positive rate = 0.7151577	True negative rate = 0.7419712

Table 2. Confusion matrix of a prediction results. Own elaboration.

As shown in the table above, the prediction is better than a fair coin toss and yields an accuracy of 72,8 % which is objectively a favorable result. It is apparent that, the accuracy of predicting decline in the share price is slightly higher than increase. This might be due to the relative absence of short selling on WSE and therefore, weak efficiency of downward moves of financial instruments prices[5].

Results of the model, however noteworthy they may be, are consistent with the results reported in literature applying similar machine learning models (e.g. artificial neural networks, support vector machines or boosted decision trees) to financial instruments on different markets (Leung M. T. et.al.; Basak S. et.al; Kumar Y.;Xinjie D.). Moreover, there seems to be

[4] Ting K.M, Encyclopedia of Machine Learning. Springer, Boston, MA, 2011, p.9.

[5] Saffi P.A.C., Sigurdson K., Price *Efficiency and Short Selling*, IESE Business school – University of Navarra, Barcelona, 2008, p.1.

no significant difference between results across examined countries.

Since returns from share prices obviously vary, it is still unsettled whether performance of a trading strategy based on proposed model will be satisfying. Thus, the author executed a Monte Carlo simulation to estimate potential performance and risk associated with investment. Stochastic process applied for a given simulation is a geometrical Brownian motion. Author assumes that the usual model for the financial asset price p_t motion is given by the mentioned geometric Brownian motion, represented by the following equation:

$$R_i = \frac{p_{i+1} - p_i}{p_i} = \varepsilon\Delta t + \sigma\xi\sqrt{\Delta t}$$

Where:

$R =$ rate of return

i = time

p = price

ε = expected value

Δ = increase in value

$\sigma =$ standard deviation

$\xi =$ random number from Laplace (μ, b) distribution, with probability density such that :

$$f(x|u,b) = \frac{1}{2b} exp - \left(\frac{|x-\mu|}{b}\right) = \frac{1}{2b}\begin{cases} exp\left(-\frac{\mu - x}{b}\right) \; if \; x < \mu \\ exp\left(-\frac{x-\mu}{b}\right) \; if \; x \geq \mu \end{cases}$$

Where:

μ = location parameter

b = scale parameter, $b > 0$

Large kurtosis or so called fat tails of stock returns distribution is the main rationale behind using Laplace instead of normal distribution. Large kurtosis is caused by the presence of black swans in the economy, i.e. events that brings highly unusual volatility on the financial markets. That is why after many such an events, economic academists criticize incorporation of normal distribution for modeling financial markets or other activities exposed to similar risk (e.g. normal distribution based VaR of bank's loan portfolio). Laplace distribution is being effectively used for that purpose, due to its fat tails and consequently, better risk assessment.

Parameters used for generating geometrical Brownian motion were derived from historical data of stock prices. Standard deviation of one trading week stock returns from past 5 years will be a scale parameter with the value of 0.039. Expected value and location were chosen so that there will be no upside or downside bias. XGboost model definitely cannot predict stochastic process, which is why it was generated with 72.7% share of positive return periods, in order to imitate predictive ability of model built earlier. Also, for the sake of realism, the process will decrease by 0.8% due to the broker's commission every period. Given assumptions described above, author generated 10 000 simulations of investments with initial capital of 1000 pln over 1 year. Chart 2 visualize executed Monte Carlo simulation.

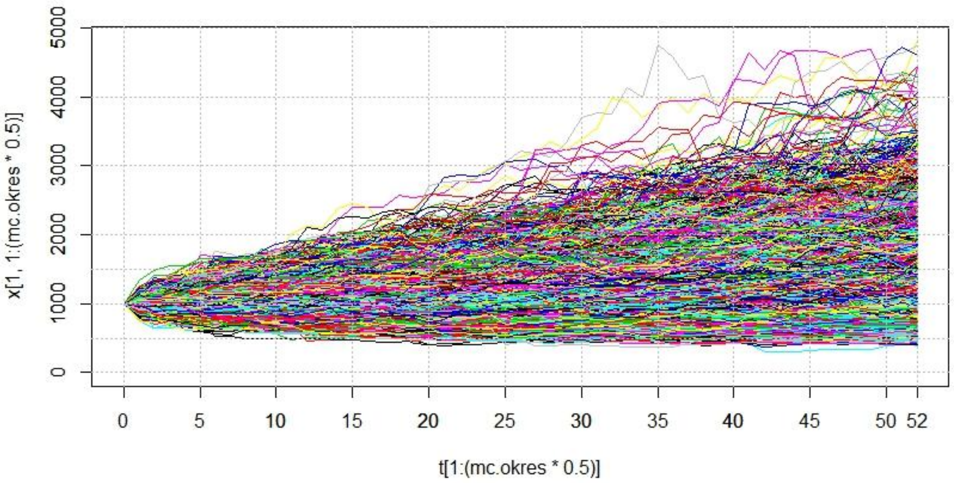

Chart 1. 10 000 simulations (in different colors) with Monte Carlo method of 1 000 pln investments with accuracy of 72.7%. X axis represents weeks and y axis represents value of given investment. Own elaboration.

The results of simulations are also satisfying. Mean return on investment of every simulation is 48.6% and the median is 39.5%, although, the volatility is also relatively high. 25% of all simulations had a return on investment less than 8,9% and in the worst case, investment lost 68% of initial capital. Chart 2 represents a histogram of final return of investment for every simulation.

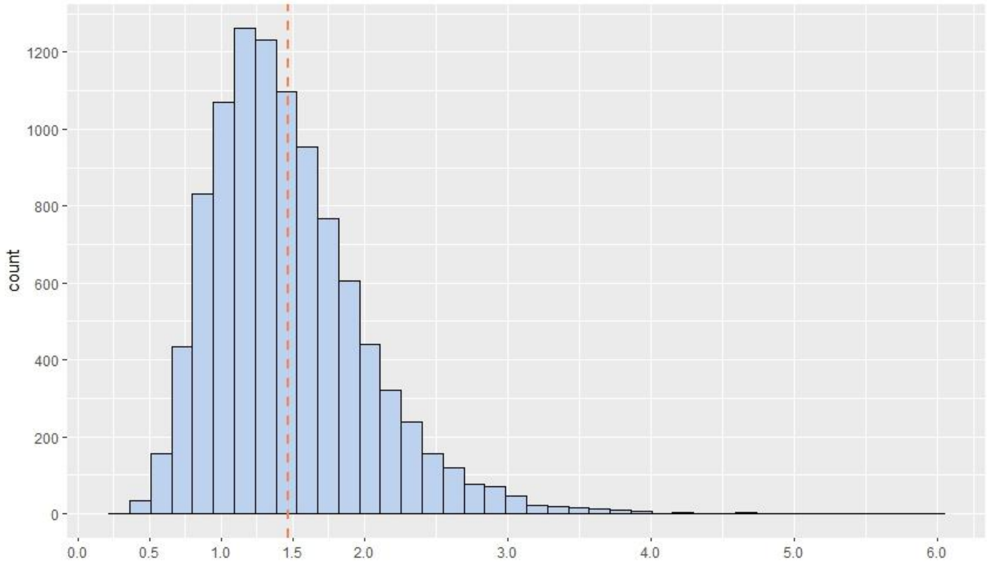

Chart 2. Histogram of return of investment for every simulation. Amount of simulations for given interval is on the y axis and ROI is on the x axis. 1 ROI is 100% of initial capital, ergo no change. Red line indicate mean. Own elaboration.

As often, backtesting trading strategy and performing simulations might look promising. Nonetheless, these, as well as xgboost model, could be flawless had they not been reliant on historical data. Considering this, there are some significant drawbacks of these methods. Notably, potential regime shift in the market microstructure, not only for particular stock but also for the whole polish stock exchange. In consequence, model learned on historical data is but a fair toss of coin, along with methods evaluating it. Results can possibly be also misleading, which is even worse. Additional defect is the potential difference in the structure of model predictive ability. It is conceivable that, for example, the model predicts small positive rate of change with better accuracy than those with higher one. In opposite, it also might be possible to correctly identify small declines more often. Apparently, the negative effect of this flaw is more probable since more volatile moves are usually effect of hard to predict real events. Another drawback is related to low liquidity of the polish stock market. Although, the chosen stock is characterized by relatively high volume, there still will be a liquidity premium involved in every transaction. Since there are 104 transactions executed in every year, this premium will almost certainly shrink the performance of investment.

Conclusion

With all the advancement in machine learning, the models from this field applied to stock market often proved to be of significant effectiveness in performing given task, whether it is classification or regression. Polish stock market also confirmed to be capable of being a suitable subject of trading with machine learning methods. While, the model, despite all of its deficiency described above, seems to be robust in predicting the stock market to some extent. Albeit, it still underperform some models with analogous algorithms presented in literature, perhaps due to different feature engineering or model tuning.

References

Basak S., et.al, *Predicting the Direction of Stock Market Price Using Tree Based Classifiers,* Applied Mathematical Finance, 2016.

Berlinger E., *Mastering R for Quantitative Finance*, Packt, Birmingham, 2015.

Chung K. L., Sahlia F., *Elementary Probability Theory: With Stochastic Processes and an Introduction to Mathematical Finance*, Springer, Boston, 2010.

Clinton V., Reddy K., *Simulating Stock Prices Using Geometric Brownian Motion: Evidence from Australian Companies,* Australasian Accounting, Business and Finance Journal, 10(3), 2016.

Fama E. A., *Capital Markets: A Review of Theory and Empirical Work,* The Journal of Finance, Vol. 25, 1969.

Focardi S. M., Fabozzi F. J., *The Mathematics of Financial Modeling and Investment Management*, Wiley, 2004.

Guestrin C., Chen T., *XGBoost: A Scalable Tree Boosting System,* Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, 2016.

Kumar Y., et.al, *Forecasting to Classification: Predicting the direction of stock market price using Xtreme Gradient Boosting, Working paper. DOI: 10.13140/RG.2.2.15294.48968.*

Laplace, P-S. . *Mémoire sur la probabilité des causes par les évènements,* Mémoires de l'Academie Royale des Sciences Presentés par Divers Savan, 6, 1774.

Leung M. T., et.al, *Application of neural networks to an emerging financial market: forecasting and trading the Taiwan Stock Index,* Computers & Operations Research 30, 2003.

Ponti M. A., *Machine Learning a Practical Approach on the Statistical Learning Theory, Springer,* Boston, 2018.

Saffi P.A.C., Sigurdson K., *Price Efficiency and Short Selling,* IESE Business school – University of Navarra, Barcelona, 2008.

Stearns B., et. al, *Schoolar Performance Prediction using Boosted Regression Trees Techniques,* ESANN 2017 proceedings, Bruges, 2017.

Stiglitz J.E., *Grossman J.S., On the impossibility of Informationally Efficient Markets*, The American Economic Review, 1980.

Ting K.M.*, Encyclopedia of Machine Learning,* Springer, Boston, 2011.

Zięba M., et.al, *Ensemble Boosted Trees with Synthetic Features Generation in Application to Bankruptcy Prediction,* Expert Systems with Applications, 2016.

Xinjie D, *Stock Trend Prediction with Technical Indicators using SVM,* Stanford University, 2014.

Streszczenie

Celem niniejszego artykułu jest wykorzystanie modelu z dziedziny uczenia maszynowego opartego na algorytmie zespołu wzmocnionych gradientowo drzew decyzyjnych do prognozowania kierunku zmian kursu akcji Banku Handlowego S.A. notowanego na GPW. We wstępie został przedstawiony kontekst uczenia maszynowego oraz wykorzystania go do prognozowania cen akcji. Następnie, przedstawiono proces tworzenia modelu klasyfikacyjnego wykorzystujący strukturę XGboost od etapu przetwarzania danych do jego ewaluacji. Danymi wejściowymi modelu były wskaźniki wykorzystywane w analizie technicznej, m.in. oscylatory stochastyczne oraz średnie ruchome, natomiast danymi wyjściowymi były kierunki zmian kursu na przestrzeni następnego tygodnia. Skuteczność modelu na danych testowych wyniosła 72%. Na końcu przeprowadzono symulacje portfela inwestycyjnego, podejmującego decyzje o transakcjach na podstawie wcześniej stworzonego modelu, wykorzystując metodę Monte Carlo w której dynamika procesów stochastycznych miała rozkład Laplace'a. Przy interpretacji wyników portfela inwestycyjnego wskazano ograniczenia ewaluacji modelu i strategii inwestycyjnej opartej o *backtest*.

Summary

The main purpose of this article is to apply machine learning model based on ensemble of gradient boosted decision trees to forecast direction of share prices of Bank Handlowy S.A listed on WSE. In the introduction, the author presented the context of machine learning and its application in forecasting stock prices. Afterwards, the author describes the process of building classification model which uses XGboost framework from data preprocessing to model evaluation. The input features of the model were technical analysis indicators, like stochastic oscillators or moving averages. Output of the model was a direction of stock price after one week. The accuracy of the model based on testing dataset is 72%. The author also performed a simulation, based on the model. The simulation was made with the Monte Carlo method which stochastic process had a Laplace distribution. During interpretation, at the end, the author pointed limitations of model and algorithmic trading strategy evaluation techniques based on backtest.