SCIENTIFIC PUBLICATIONS OF THE STATE UNIVERSITY OF NOVI PAZAR
SER. A: APPL. MATH. INFORM. AND MECH. vol. 11, 1 (2019), 47-54.

Machine Learning Techniques for Stock Market Trends Identification

Ekaterina Zolotareva

Abstract: The research concentrates on recognizing stock markets long-term upward and downward trends. The key results are obtained with the use of gradient boosting algorithms, XGBoost in particular. The raw data is represented by time series with basic stock market quotes with periods labelled by experts as Trend or Flat. The features are then obtained via various data transformations, aiming to catch implicit factors resulting in change of stock direction. Modelling is done in two stages: stage one aims to detect endpoints of tendencies (i.e. "sliding windows"), stage two recognizes the tendency itself inside the window. The research addresses such issues as imbalanced datasets and contradicting labels, as well as the need of specific quality metrics to keep up with practical applicability. The model can be used to design an investment strategy though further research in feature engineering and fine calibration is required.

1 Introduction

An ability to identify stock market trends has obvious advantages for investors. Buying stock on upward trend (as well as selling it in case of downward movement) results in profit. With the rise of machine learning in early 2010s researchers started to take interest in applying computer science to financial market problems [2]. Some market experts still argue that traders are able to see opportunities of making profit (i.e. detecting trends or turning points) which can not be formally expressed. Thus using computer science algorithms to learn from successful traders' decisions (and not only stock data) is likely to improve financial market models.

Manuscript received July 20, 2018; accepted February 15, 2019.

Ekaterina Zolotareva is with the Department Financial University under the Government of the Russian Federation 38 Shcherbakovskaya St., Moscow 105187, Russia

2 Problem formulation

The research was conducted in behalf of one major Russian investment company (The Company). The Company experts have labelled historical S&P stock data, that is they marked certain consequent periods as "Trend", "Flat" or N/A in a specially designed software with a graphical interface. Approximately 90% of identified trends last between 40 to 600 business days, which accounts for middle- or long-term tendencies. Initially the task was to train the model to identify the trend itself (no matter the direction) with the minimum lag from its start. Later, though, it turned out that it is also necessary to distinguish between downward or upward trends in order to calculate and compare the financial results of different strategies. The model should be independent from any specific stock, market or time period, after it is properly calibrated it should be equally good for any asset and time. Another important issue is that by learning from historical patterns we aim to identify the current market situation (answer the question: what long-term tendency takes place today?) and we must always bear in mind that future data is unavailable. Breaking this condition will make the modelling results irrelevant, though minor time lag (within a couple of weeks) in quite acceptable.

3 Data overview

The dataset to explore consists of two sources (Source I and Source II), labelled by 9 experts. The data contains quotes of 705 stocks covering 705 stocks for the period from 2005-01-28 to 2017-09-13. The sources have an intersection in time period - dates from 2007-08-08 to 2017-05-24, but they have only one intersection in the list of stocks. Only 4 experts labelled both Source I and Source II data, but the second dataset in considered "cleaner" since the experts were more motivated to label data responsibly. The total number of records in both datasets amounts to 9 180 712 pieces packed in 3162 files. Each file contains on average around 2600 daily quotes (Date, Open, High, Low, Close) for a certain period and stockname, labelled by a certain expert.

4 Model outline

Modelling is done in two stages: stage one aims to detect endpoints of tendencies ("change points", or "turning points"), stage two recognizes the tendency itself inside the window.

The performance of stage one is provided by the model which will be referred to as "ChangePoints". For each data point it returns either a value "1" ("The change of tendency occurred" or "A new window has started") or 0 ("No changepoint"). Due to various reasons, discussed later, currently the ChangePoints predictions are subject to both false positive (mainly) and false negative errors. This is why it can't be used alone and should be backed by the stage two model -"TrendOrFlat". TrendOrFlat is launched when the possible start

of new tendency is detected, i.e. “Change points” returns “1”. It is important to note that once the changepoint signal occurs, we come to recognize the starting point of the new tendency, but we do not know how long it will last or when the endpoint occurs. To identify the tendency inside the new window TrendOrFlat model initially analyzes the first few, say 6, days of it, then first 7 days, then 8, etc., returning the values “1” (“Upward trend”), “-1” (“Downward trend”) or “0” (“No trend/Flat”) for each period. Shortly after the start , TrendOrFlat is more likely to produce incorrect predictions, but as the window widens, the tendency identification becomes more and more accurate. The process continues up until a new positive signal from ChangePoints occurs, indicating the start of a new window for which the routine repeats. The final results of modelling are determined upon TrendOrFlat output. The whole process, run for a set of instruments (stocks) on the chosen time period, will be referred to as ‘pipeline’. The calculations were processed on Python 3.5.

5 Train and test sample

In order to evaluate the generalizing power of the model we traditionally divide the dataset into train and test samples. These samples should be independent, otherwise the quality metrics would be misleadingly inflated. Our final choice is to use 70% of older data as a train set and the remaining 30% as a test set. The breakpoint date is October 14^{th} 2014 if both Source I and Source II were analyzed and November 6^{th} 2014 if only Source II was used.

6 ChangePoints model

We start from the ChangePoints model and its features. It so happened that this model is subject to various complicated issues, while TrendOrFlat works fairly smoothly. The list of ChangePoints features is presented in Table 1, totaling 22 input variable plus 1 target.

First the raw features were used for modelling. Later it appeared more appropriate to use natural logarithms instead, since they are less subject to the spreads in absolute values. Furthermore, it turned out that experts were using logarithmic price scales when labelling the data.

7 Contradicting labels issue

Each expert would label a certain data point only once, but different experts can label same data points and their expertise does not necessarily coincide. This results in several thousands of records with identical input features but different target values. Literally every positive record has a negative contradict. In order to suppress this issue the following strategies have been introduced:

Table 1.

Features	Description	
	Raw	Logarithmic
Close-1, Close-2,... Close-5	Ratios of the 5 previous closing prices (today minus 1, minus 2 and so on) to the current closing price.	The natural logarithms of the corresponding ratios, or the difference between the logarithms of numerator and denominator
Close1, Close2,... Close5	Ratios of the 5 future closing prices (today plus 1, plus 2 and so on) to the current closing price.	
Volume-1, Volume-2,... Volume-5	Ratios of the 5 previous trading volumes (today minus 1, minus 2 and so on) to the current trading volume.	
Volume1, Volume2,... Volume5	Ratios of the 5 future trading volumes (today plus 1, plus 2 and so on) to the current trading volume.	
High	Ratio of today's maximum price to the closing price	
Low	Ratio of today's minimum price to the closing price	
NewTrigger	Target variable, indicating whether the change in tendency has occurred today or not. It takes value "1" the day the tendency changes and remains "0" otherwise.	

1. Averaging the experts opinions, or voting.
2. Triggers correction (this alternation to data targets the technical blot issue).
3. Excluding irrelevant experts.
4. Ignoring the contradictions. The major pitfall of these approach is that traditional classification quality metrics (Accuracy, AUC, Precision, Recall, F-Score) will become irrelevant, since there is no "ground truth" anymore [1].

8 Imbalanced dataset issue

From a formal point of view, ChangePoints model is a binary classification model. In a perfect situation the proportion between classes should be close to 1:1, otherwise the observations of a minority class would be "surpressed" by majority. The traditional classification

quality metrics, on the contrary, would perform quite well, unless you drill down to contingency matrix or evaluate performance on the minority and majority classes separately. Unfortunately, our dataset is highly imbalanced: records where "NewTrigger" equals "1" (positives) are the minority class with the proportion varying from 78:1 or even 331:1 depending on source filters and the application of averaging. That is natural because we seek for changepoints of middle-and long-term tendencies which would happen only once in several hundred business days.

9 Quality evaluation issue

As we can see, there are at least two important groundings against traditional classification quality metrics (Accuracy, AUC, Precision, Recall, F-Score): they can be misleading both due to "ground truth" contradictions and highly imbalanced dataset. But we will find another reason to consider them irrelevant if we remember the time series issue. In a common classification model the records are absolutely independent and if we shift the prediction of "1" to the neighboring record, we'll have a completely different result. But with the ChangePoints model, shifting one day back or forward will result in only a very minor change in terms of profit. This brings us to the conclusion that the most relevant quality metrics for such kind of models are those profit-related.

10 ChangePoints XGBoost realization

There are a number of classification algorithms –e.g. logistic regression, Bayesian classifier, SVM, neural networks, decision trees and their ensembles. Out of all these diversity, it is gradient boosting ensembles which account for best modeling results during the last couple of years. The modelling in this research was completed with XGBoost [3]. Another advantage of the XGBoost algorithm is that it allows to control the imbalanced dataset issue by directly setting the hyperparameter "scale_pos_weight" to the proportion between the negative (majority) and positive (minority) classes in the dataset ("balance"), ensuring the parity between the classes while training the model.

The best hyperparameters were searched by grid search/randomized grid search procedures, but after a number of iterations it became obvious that it is mainly 4 hyperparameters that matter and even they do not alter the result crucially. The characteristics of the ChangePoints model, which demonstrated the best result on the pipeline are the following: n_estimators=500, max_depth= 7, reg_lambda=3, subsample=1, learning rate=0.1, scale_pos_weight=154, seed=42, nthread=-1, other parameters set to default. The train set of the best model contained logarithmic data from both sources, but only two experts out of 9 were left. These experts were the main stakeholders of the research, with greater experience and motivation. No averaging of experts opinions was applied, though trigger

correction was.

The overall performance of the model is quite good with AUC=86.07% and F-Score= 95% on test set. However, drilling down we can see, that the performance on the minority subset is quite poor (F-score=8%), mainly due to the extremely low precision (5%).This means our model creates too much "false alarms", detecting non-existent changepoints. The recall value for the minority subset is 58%, meaning that a vast amount of true changepoints is also missed.

11 TrendOrFlat model

Table 2 gives the list of suggested features (5+1 target) and the intuition behind them.

Table 2.

Feature	Description
RegClose	The slope of the linear regression line for daily closing prices (logarithmic or not).
CloseR2	The R2 coefficient of the linear regression line for daily closing prices (logarithmic or not).
RegVol	The slope of the linear regression line for daily volumes (logarithmic or not).
VolR2	The R2 coefficient of the linear regression line for daily volumes (logarithmic or not).
LenTrend	The length of tendency (or the width of window) in business days.
NewTypeBool	Target variable. The Boolean analogue of the "Type" field. It takes the value "1" in case of trend and "0" otherwise.

12 TrendOrFlat dataset overview

The dataset for the TrendOrFlat model is different from that for the ChangePoints, since now we are dealing with time periods, not separate trading days. The size of the dataset equals the total number of windows marked by experts.

It is important for the model to recognize tendencies by their parts. To ensure this we supplemented the initial dataset, which contained full windows only, by feature vectors extracted from 5, 10, 20,. . . .90% parts of full windows. The total number of records in the train set varies from 97 225 to 232 870 depending on the source filters, with 10 to 20 thousands of full windows. The sample is balanced and does not contain contradictions.

13 TrendOrFlat XGBoost realization

TrendOrFlat model is also a binary classification model. We used the XGBoost algorithm again with the following set of hyperparameters: n_estimators=100, max_depth=5,reg_lambda=3, learning_rate=0.2, seed=42, nthread=-1, others set to default [3]. Again the train set for the best model was based on logarithmic data from both sources, but only two best experts were left.

The best scores are AUC= 81.86% and F-score =74% and, as expected, they do not vary within classes. Even for very small parts of trends the classification quality is around 70% and sufficiently increases up to 90-95% when 80% days or more are shown to the model. That means that TrendOr Flat model is likely to correct ChangePoints pitfalls. Also note, that if the output of the model is 1("Trend") we can easily determine the trend direction from the RegClose value (positive for upward trends and negative for downward).

14 Pipeline results

The pipeline logic is described in the beginning of the paper. Here we shall concentrate on the discussion of the specific quality metrics. Recalling that the direction of trend can be determined by the slope of the regression line, we can calculate the profit earned during the time in position (i.e. during trends) and a couple of other metrics, as presented in Table 3.

Table 3.

Indicator	Description
Profit	The sum of all profits earned during the time in position (for all the stocks).
Days_in	The total number of business days in position (for all stocks)
DayProfit	The profit per one day in position,% : DayProfit= Profit/ Days_in
YearProfit	dayProfit scaled per annum,%: YearProfit= DayProfit*250, where 250 is the average number of business days in a year
YearProfit_avg	The average annual profit, including the days not in position, %: YearProfit_avg = Profit/number of datapoints in the dataset.

For the purpose of comparing models, the last two indicators - *YearProfit* and *YearProfit_avg* - are the most informative, because they are independent from the length of time period and the number of stocks in pipeline. From a business point of view it might also be important

to see how many times we opened the positions or what was the proportion between short or long for each stock, because it influences the additional costs of trading.

The best results of all the pipelines tested are the following: *YearProfit =28.8% and YearProfit_avg=6.9%*. Of course, they are much more modest than the ground truth –the experts who were dealing with historical data, or "saw the future", achieved at least twice more. Nevetherless, having nearly 30% on investment per annum looks quite impressing. Unfortunately, the model still did miss a lot of opportunities (76% of records are predicted as flat), so *YearProfit_avg* is not too big. Though If we add overnight interest paid on flat periods (say 7%) our average profit will reach 7%*0.76+6,9%=12,2% per annum.

15 Conclusion

The model, presented in the research, can be used by both individual and institutional investors. It produces "buy" and "sell" signals when starting or endpoints of trends are identified. The profit earned on days in position can reach 28.8% per annum, but definitely the result can be improved.

There are several directions for this work:

1. Implementing other approaches to deal with contradicting labels and the imbalanced datasets – two major issues which influence the quality of the ChangePoints model.

2. Selecting other sets of features for the ChangePoints and TrendOrFlat models. Various combinations of technical indicators should be tried for ChangePoints models and probably different time lags and threshold levels. As for TrendOrFlat, the improvement should be concentrated on early tendency identification.

3. And finally, totally changing the model structure and using other machine learning algorithms, for example, convolutional neural networks, can also sufficiently improve the model.

References

[1] H. BRINK, J. RICHARDS, M. FETHEROLF, *Real-World Machine Learning*. Manning Publications Co., USA, 2017.

[2] V. SOLOVIEV (2017) *Forecasting Stock Market Turnovers with Boosted Decision Trees* Proc. 11th International Conference on Application of Information and Communication Technologies (AICT) 2017, Vol. 1, pp. 140-143.

[3] XGBoost Parameters.
http://xgboost.readthedocs.io/en/latest/parameter.html