General Bayesian Networks for Stock Market Analysis

Camila Ozelame [1], **Rodolfo Ferrari** [2], **Leando Mundim** [3]

1 Introduction

Data science and trading algorithms are improving each day and the spotlight is not only on the best guidance of traders, but a robust guidance backed by a reliable mathematical-statistical algorithm. Understanding trends in the market is always a good indicator of how a stock or an index can behave following investors reactions. Statistics tools such as Bayesian Networks on a complex environment of equities tradings influenced not only by internal financial ratios and forces but also macro and microeconomics could represent a strong mechanism for a better understanding and helping traders to take decision more data-oriented rather than only relying on their sentiments and believes on market behaviour. Furthermore, some forces such as LIBOR, Unemployment Rate or even Consumer Price Index can be explained by any economist on an isolated manner impacting a stock or an index, but the questions are: how those factors added to intrinsic returns of any stock behave daily on a complex network? and how the network can mutate as long as time pass by?

2 Methodology

The main purpose of this work is the daily prediction daily tendency of *each stock* based on Sliding Window and Bayesian Networks Analysis. The tendency will be measured according to the return value, which is further described, indicating the best action given the current scenario.

2.1 Literature Review

2.1.1 Briefing about Bayesian Networks

Bayesian Networks are, basically, a visual representation of a joint distribution of a group of random variables in a **D**irected **A**cyclic **G**raph (DAG), it was first proposed by Judea Pearl. The variables are represented by nodes and the conditional dependence between them are represented by arcs. By definition, the joint distribution is the product of the probability of each node given its parents, represented by the equation below:

$$P(\mathbf{X}|\mathbb{G}) = \prod_{i=1}^{d} P(X_i|\mathbf{pa}_{\mathbb{G}}(X_i))$$

Where $\mathbb{G}$ is the graph structure Bielza e Larrañaga (2014). In Figure 1 we have an example of a Bayesian Network, where Y depends on the set $\{X_1,...,X_5\}$, that is the variables in the set $\{X_1,...,X_5\}$ are the parents of Y. In the same way, X_6 and X_7 are the parents of X_2 and X_8 is the parent of X_4. The lonely node X_9 does not have any connection with the network so we say that it is independent of the rest of the variables.

[1] PIPGEs - UFSCar/USP. e-mail: *ozelamecs@usp.br*

[2] PPG CCMC - Universidade de São Paulo. e-mail: *rodolfoferrari28@gmail.com*

[3] ICMC - Universidade de São Paulo. e-mail: *leandroresendemundim@gmail.com*

Some proprieties of graphs allow the BN to be an useful tool for the tasks of modeling the connections between variables, reducing the data dimension, allowing better interpretation of the data and also prediction.

Several classifiers based on Bayesian Networks are available in literature and for the specific application, the *General Bayesian Network (GBN)* was chosen for the special reason that it does not make any assumption about the relationship among variables, it uses techniques for the structure estimation considering the class variable as an ordinary variable and then, calculates the probabilities for each possible response.

For the structure estimation, some heuristics, such as score-based and rule-based (also known as constraint-based algorithms), are used to tackle the problem and also, hybrid techniques can be found in literature. In this study, we focus on the score-based method called *Hill Climbing*. This procedure starts from a specific structure in space (randomly chosen or based on prior knowledge) and considers all neighboring structures obtained from the current structure by adding, deleting, or reversing a single edge at every iteration of the search algorithm.

The search progresses to the neighboring structure having the highest value of a score if this value is higher than that of the current structure and it stops when reaches a local maximum Lerner e Malka* (2011). The score, in this method, makes the role of quantifying the amount of information that structure represents, we choose the k2 score which is widely used in literature for classification tasks.

2.1.2 Sliding Window

Another technique present in this paper is the *Sliding Window* that is widely used for adaptive filtering also detection models Weinberg (2017). It can be adapted for many approaches but here it is considered as a model based on recent data Bodenham (2012). This method sets a window of a fixed size L and the model considers only the more recent tuples in the data, where $(\mathbf{X}^i, Y^i)$ represent the set of explain variables $\mathbf{X}$, the class variable Y at the time i. The temporal data is represented bellow:

$$Data : (\mathbf{X}^1, Y^1), (\mathbf{X}^2, Y^2), \ldots, (\mathbf{X}^{t-L}, Y^{t-L}), (\mathbf{X}^{t-L+1}, Y^{t-L+1}), \ldots, (\mathbf{X}^t, Y^t)$$

The window slides according to the tuple index we are predicting, for instance, when the goal is to predict the Y^{t+1} the training subset is: $\{(\mathbf{X}^{t-L}, Y^{t-L}),$
$(\mathbf{X}^{t-L+1}, Y^{t-L+1}), \ldots, (\mathbf{X}^t, Y^t)\}$.

The method described in the subsection 2.1.1 is a non-temporal Bayesian Network, the temporality is included in the model by the Sliding Window technique, subsection 2.1.2 which considers the most recent data to train the network.

2.2 Proposed Approach

In this section we present the variables will be used in the study that are both, provided and calculated based on the data. And also, we discuss about the selection of periods of time, discretization methods and the analysis methods.

2.2.1 Available Information

In the Table 1 the relevant variables - an expert point of view - considered in the model are listed and in another work they are deeply explained, however an important description of them is saying that they can be labeled as price variables, technical analysis and momentum index and external factors.

Table 1: Description of the variables involved in the training model.

Variable	Description	Variable	Description
Y	Class Variable	P_BR	Price-to-Book Ratio
volume	Number of Shares Traded	P_SR	Price-to-Sales Ratio
VIX	Implied Volatility Index	EPS	Earnings per Share of S&P500
LIBOR	Interest Rate	RSI	Relative Strength Index
CPI	Consumer Price Index	dn	The lower Bollinger Band
UR	Unemployment Rate	mavg	Moving Average
HS	Housing Start	up	The upper Bollinger Band
MS	Money Supply	pctB	The %B calculation
PE_ratio	Price-to-Earnings Ratio	tdi	Trend Detection Index

2.2.2 Pre-training Parameters

This parameters indicate the best trade off between test and train subsets proportion once the test is composed only for the day that will be predicted towards to minimize the risk of a prediction method (function). This training subset is the L-sized sliding window and the preliminary results indicated that this period may vary between 110 and 115 days then, we chose an *112*-day window to train the model. For the variables mentioned before, which use periods of time to be calculated, such as Moving Average, the range selected is 14 days (which is also considered to calculate the RSI).

2.2.3 Discretization

The discretization consists in a transformation of the values in categories and it is required for the current method. The target variable is "return" that it is discretized in "sell", "hold" and "buy" according to $median \pm \frac{1}{2} sd$ - trying to avoid the outliers influence and using the historical data available. Other variables will be discretized according to its definition, e.g. RSI claims the categories of "overbought", "intermediate" and "oversold"; the ones with no such clues, the dichotomization Rucker, McShane e Preacher (2015) given the variable change (daily change compared with the previous day).

2.2.4 Performance Evaluation

The performance is evaluated as it follows, after the pre-processing phase, the data is divided into training and test, the training subset is composed for the L most recent instances and the test is the data for the "next day".

- Accuracy (ACC) which is the proportion of correctness, $1-risk$, its range is between 0 and 1;
- Matthew's Correlation Coefficient (MCC) measures the general classification of the model and the interpretation is similar with the Pearson correlation coefficient Louzada e Ara (2012), it can vary between -1 and 1;

3 Results

The study has been conducted using the software `R`, with the interface `R-Studio`. The main packages used to this tasks are `bnlearn` - to learn the graph structures and their parameters, `ggplot2` - to construct the visualization for the graph and `TTR` - Technical Trading Rules to construct indexes and functions for the analysis.

3.1 Pre-Processing

In section 2.2, a description for the variables aggregated to the model is provided and also for important parameters for the indexes construction. Some of the variables are available in the Kaggle website, the majority of the indexes are calculated in R and the external factors are found at the Bloomberg platform. The discretization for the explain variables follows.

In the package `TTR`, there are functions to calculate the Bollinger Bands, Relative Strength Index, Trend Detection Index and return, well set with the parameters such as *n*, the number of days taken into account for the indexes calculation. The literature recommends $n = \mathit{14}$ for RSI, $n = \mathit{20}$ for the others; the cut points are available for the RSI which is 30 and 70 for *overbought*, *intermediate* and *oversold* and TDI has the numbers 20 and 80 for the same categories.

The index with no such established cutoffs are categorized as *increasing*, if the value of the current day is superior comparing with the previous day and *decreasing* for the opposite, the change value is the criteria.

The return categorization is described in the subsection 2.2.3. After the discretization, the database is subset in $L + \mathit{1}$ days and the trained model is used for the daily prediction. The tests were conducted using the stock "AAPL", for the balance of the response variable and also indicated some insights about the problem, such as the window size of 112 days for the chosen method and score.

3.2 An Overview

Once the windows size has been chosen, $L = 112$, we estimated the accuracy and the Matthew's Correlation Coefficient for all stocks in the database. In order to find reliable results about the method analysis, stocks with less than 1258 observations and/or with missing values were not considered. In this case, the number of stocks in the analysis is 453.

The rankings for the best values of the performance measures are presented in Figure 1 until the 20^{th} position. Notice the axis do not have same scales and limits also, the shares are different in each chart.

Lerner, in Lerner e Malka* (2011), aims the K2 algorithm performs better the task of inference rather than prediction. A visualization of one prediction for the stock "AAPL" is presented in Figure 2. Furthermore, the technique provides the probabilities to buy (0.0958), **hold (0.8944)**, sell (0.0098), the true class is "Hold", with P_BR (Price-to-Book) as a parent.

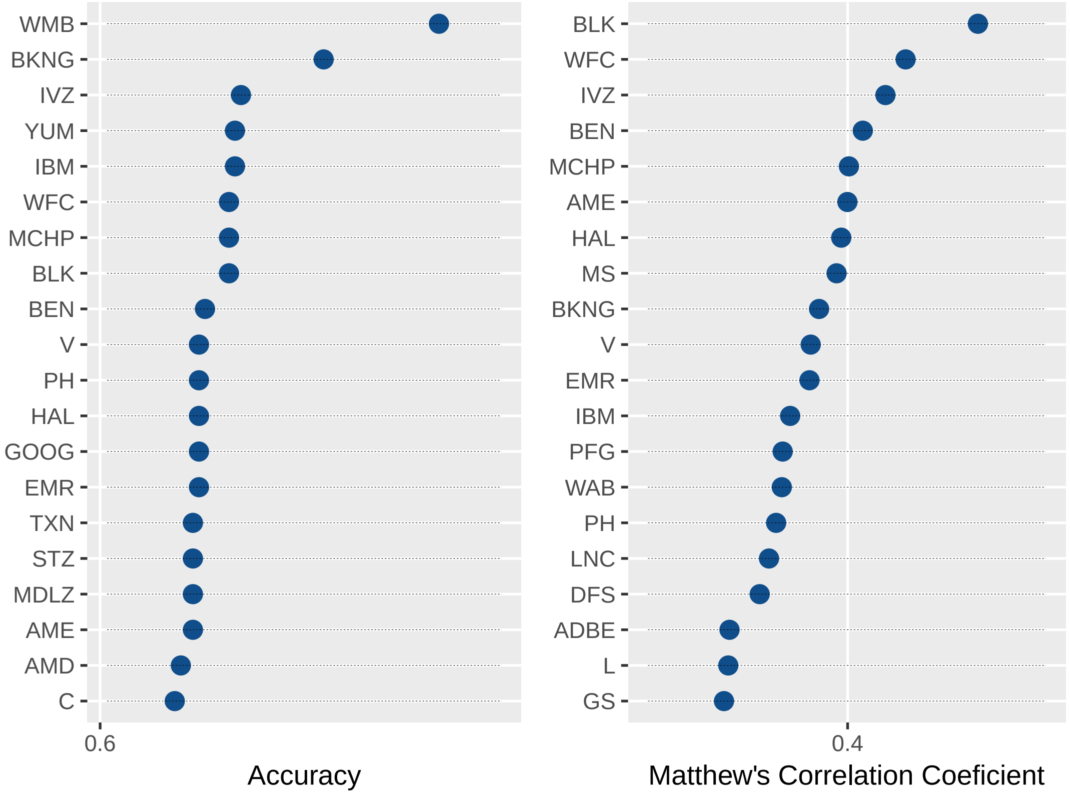

Figure 1: Highest values for the performance measures.

4 Discussion

This work focus on modeling the S&P500 using Bayesian Networks and Sliding Window techniques to tackle the problem of predicting the the daily tendency of the stocks; this tendency is calculated based on the daily return and classified as "buy", "hold" and "sell", all based on the historical data.

The analysis returns an average accuracy of 0.6 and the standard deviation of 0.02, while Matthew's Coefficient has the average of 0.37 and the standard deviation of 0.05.

For a more specific study, we selected one stock to study the interference of the variable in the class response, it is observed the variables that are most connected with the answer as long as the time goes. The technique provides the model using the 112 most recent observation, gathering the patterns and estimating the most probable class tendency for the day.

Furthermore, the graph also works as a variable selection method, once the nodes that are not connected with the whole network can be considered independent of the structure. However, in the period of investigation, the lack of edges is given by the non observation of the variable level, for instance, CPI usually has two levels - "increasing" and "decreasing" - in this case, one of them did not appear in the time window considered for the training.

5 Conclusion

The combination of techniques presented in this work gives a powerful tool to guide traders on stock market daily decisions, providing probabilities of the tendency and also a visual guide for the influencers of the interest variable, basically, it joints the inference

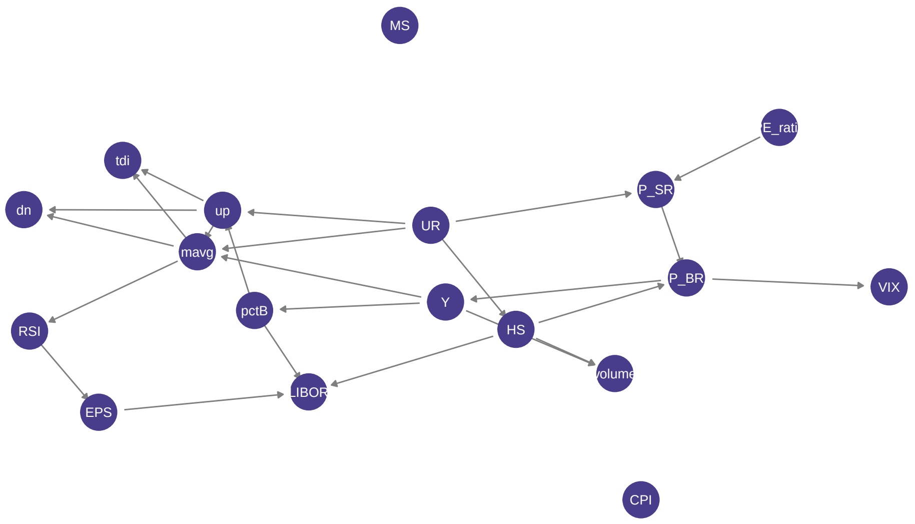

Figure 2: Generated graph to predict the "AAPL" in "2017-03-07".

with the predictions. The tests show, on specific points, the closest variables that affect the network result and how it changes as long as the time passes by also, we are able to infer about the variables relationships and the set of nodes that, directly or not, impacts the prediction.

Acknowledgements

This study was financed in part by the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES)* - Finance Code 001.

References

BIELZA, C.; LARRAÑAGA, P. Discrete bayesian network classifiers: a survey. *ACM Computing Surveys (CSUR)*, ACM, v. 47, n. 1, p. 5, 2014.

BODENHAM, D. *Adaptive filtering and change detection for streaming data.* Tese (Doutorado) — PH. D. Thesis. Imperial College, London, 2012.

LERNER, B.; MALKA*, R. Investigation of the k2 algorithm in learning bayesian network classifiers. *Applied Artificial Intelligence*, Taylor & Francis, v. 25, n. 1, p. 74–96, 2011.

LOUZADA, F.; ARA, A. Bagging k-dependence probabilistic networks: An alternative powerful fraud detection tool. *Expert Systems with Applications*, Elsevier, v. 39, n. 14, p. 11583–11592, 2012.

RUCKER, D. D.; MCSHANE, B. B.; PREACHER, K. J. A researcher's guide to regression, discretization, and median splits of continuous variables. *Journal of Consumer Psychology*, Elsevier, v. 25, n. 4, p. 666–678, 2015.

WEINBERG, G. *Radar detection theory of sliding window processes.* [S.l.]: CRC Press, 2017.