Analyzing stock market data values with soft sets decision making model

Hubert Lechowicz, Kamil Pierzchała
Faculty of Applied Mathematics
Silesian University of Technology
Gliwice, Poland
email: hubelec784@student.polsl.pl

***Abstract*—Stock trading - it has always been attractive to many, mostly due to the money involved. This issue is concerned about making the most profitable decision on the stock market, under the careful guidance of written program. It is not a blind guess, nor a gamble, it is rather a mathematical model that analyses the stock values using soft sets. It simply requires just the stock values from previous weeks, so the model can analyse them, and return a viable approach to the user. Stock analysis is a necessity for all brokers and other stock market players. The most popular approach to stock analysis were groups of brokers using their expert knowledge and experience. Although stock market is open for everyone willing to invest, it's hard to predict what will happen the day after investment. With soft sets, expert knowledge can be defined and aggregated into a mathematical model presenting the data in a more 'Human' outcome, opening the stock market for everyone.**

I. Introduction

A theory of decision making processes is mostly based on artificial intelligence models. Data on the input is divided into classes among which comparisons are made. Classifiers and prediction methods are very often done by the use of neural networks. Where very often probabilistic models are used as predictors [1], while also hybrid compositions with other artificial intelligence are used ie. for system positioning [2]. In recent year very high popularity in data science is achieved by soft sets. These are models of decision making systems where rules over data samples are related to the data structure. One of the first approaches to define a soft set theory was proposed in [3]. This idea started many interesting research both on theoretical aspects and practical applications. In [4] was presented how to use a soft set theory for group compositions. Article [5] presented a new composition of this approach to extend the idea of fuzzy sets. Now the theory of sots sets is composed with probabilistic approaches to strengthen decision making processes [6]. Among important applications of soft sets we can present medical processes evaluation [7].

In this article we have used a theory of soft sets to compose a market data classifier. In our approach a data from market stock is collected to compose a relation table over which we have used our soft set model to help on sell/buy decisions. The results show that our method works well and proposed classifier can be an interesting help in economic processes.

II. Model composition

Let us now discuss how we have developed our solution. In this section we will present the data used for our research and the decision making model in which a soft set approach was implemented.

A. Data preparation

All of the data needs to be prepared for calculations. Chronological order of entries is required, as it is the most important rule. In the case when it is not obeyed, methods used and described below won't be able to make any sensible prediction, due to their very nature.

B. Pandas

It is an open source, python library which provides well-optimized, data structures for the data analysis in python programming language. It is used to envelop data in data-frame structure. The web repository of this library is present at pandas.pydata.org/pandas-docs/. Implemented solution works in according to flowcharts presented in figure 1.

C. Expert knowledge

Set of rules which derivatives from actual knowledge on the subject. Usually, it is defined by an expert in the relevant field. Here it is based on through read of financial statement analysis books and guides. Rules are used to evaluate the current state in the current moment. It is not a mathematical algorithm used for calculations of some sort, but an evaluation of the output in decision model.

D. Decision model

Composed of the expert's knowledge rules, made into simple mathematical equations. Deliberately it's intended to provide a suitable prediction of the most profitable decision. Three different elements are joint together, their aggregated output is the one that is the actual value, on which prediction is based. These elements, are: Stochastic Oscillator, Moving Average Convergence / Divergence, and Relative Strength Index. Mostly refereed to, by shortened names: SO, MACD, RSI.

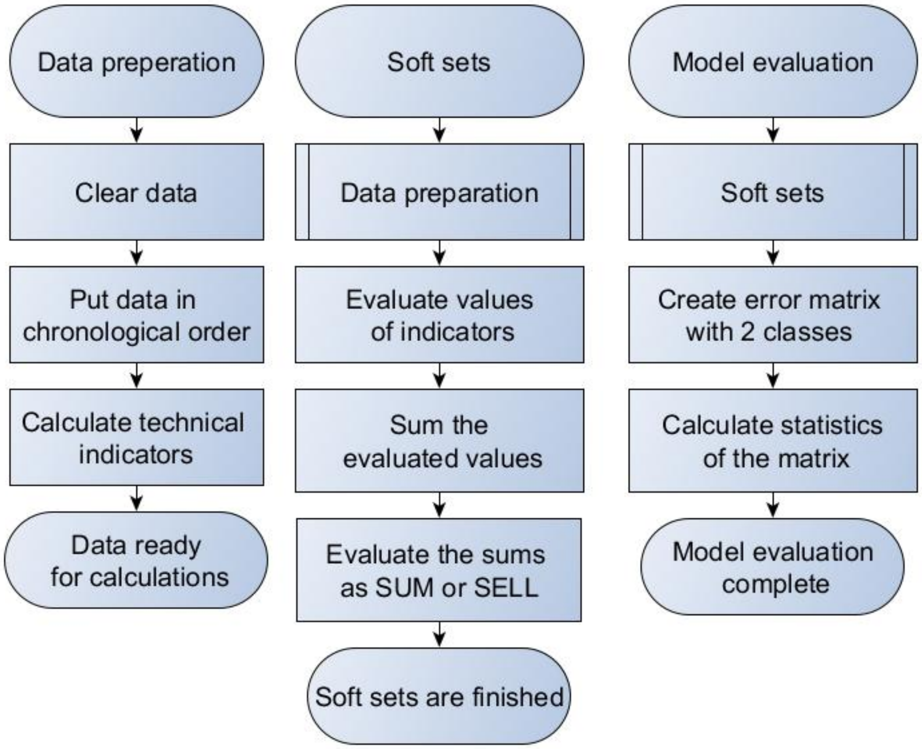

Data preperation
Clear data
Put data in chronological order
Calculate technical indicators
Data ready for calculations
Soft sets
Data preparation
Evaluate values of indicators
Sum the evaluated values
Evaluate the sums as SUM or SELL
Soft sets are finished
Model evaluation
Soft sets
Create error matrix with 2 classes
Calculate statistics of the matrix
Model evaluation complete

Figure 1: Sample flowchart of our implemented decision making model.

E. Confusion matrix (Error matrix)

Machine learning and especially the sophisticated issue of output classification has come up with a confusion matrix, more often, called error matrix. It is a definite table layout which allows visualization of an algorithm's performance. Each and every matrix row represents the predicted class instances while each and every column represents the actual class instances. It's called like that, due to the matrix making it easy to see whether the system is confusing two classes or not. That being said, the simplest description would be a genuine type of contingency table, that consists of just two dimensions ("actual" and "predicted"), and very same sets of "classes" in both dimensions.

F. Soft sets

Simplification of fuzzy set, used to when working on data with uncertainty in a parametric manner. The mathematical description of it, and simplified model is described in the next section.

III. Mathematical description

Soft sets: Simple definition of soft sets is (F, A) is named a soft set over U, where F is a described as:

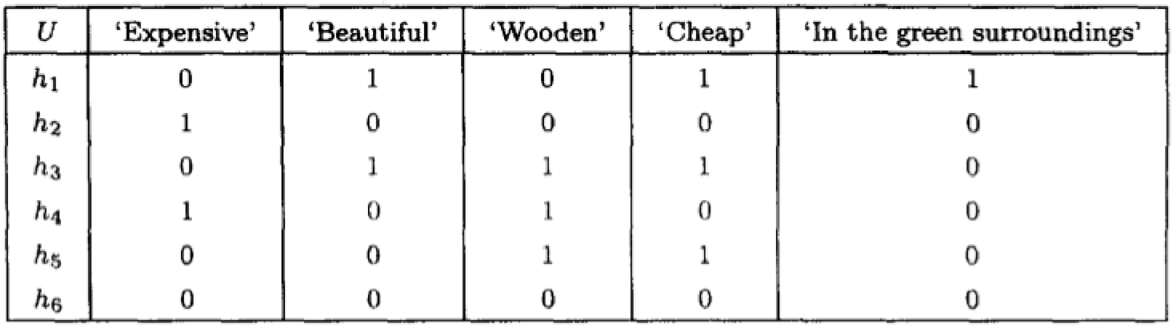
$$F : A \to P(U) \tag{1}$$

That being said, a soft set over U is a family of subsets of the universe U. For $\epsilon \in A$ F(ϵ) may be considered as the set of c-approximate elements of the soft set (F, A)

Figure 2 describes a following scenario. U is the group of real estates taken under consideration. E is the group of parameters. Each and every single element of E is a word or a sentence. E = expensive; beautiful; wooden; cheap; in the green surroundings;. Therefore, defining a soft set would be pointing out expensive estates, beautiful estates, and so on. The soft set (F, E) describes the “attractiveness of the estates” to the general customer. Shown below, soft sets are used as presented in Figure 3. They perform the role of the actual decision model. Aggregating the values of the *MACDboolean*, *SOboolean*, *RSIboolean* outcomes in the *Sum* column. The decision is listed in the *Decision* column. These three columns have Boolean in their names, as they take only three different values $-1, 0, 1.$ or just $-1, 1.$ Theoretically speaking, these values are not strictly Boolean, but since they represent positive state, negative state and unknown state, it was shortened to positive state, negativity of positive state and unknown state. Which sums up to just two values, just like Boolean values.

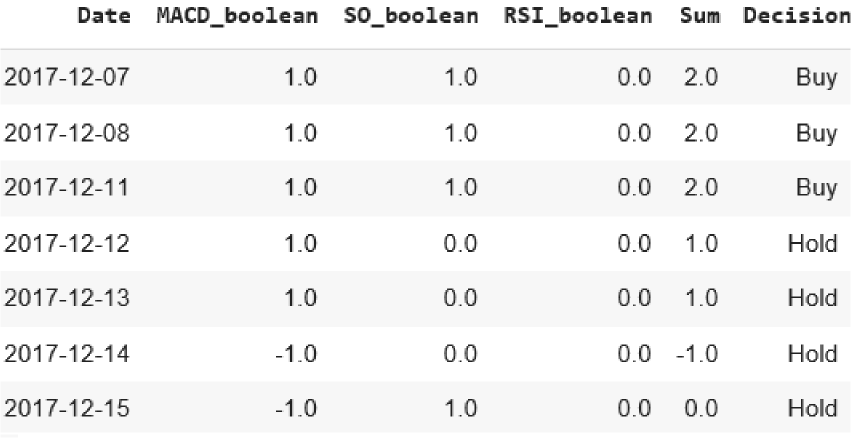
Relative Strength Index (RSI):

Technical indicator made to show the ups and downs of a stock, based on the closing prices of an n window length. It is a momentum oscillator, describing the magnitude and velocity of price movements. Momentum as rate of the rise or fall in price. RSI calculates momentum as the $highercloses/lowercloses$. The stronger the positive changes the higher RSI is. The stronger the negative changes the lower RSI is. The RSI



U	'Expensive'	'Beautiful'	'Wooden'	'Cheap'	'In the green surroundings'
h_1	0	1	0	1	1
h_2	1	0	0	0	0
h_3	0	1	1	1	0
h_4	1	0	1	0	0
h_5	0	0	1	1	0
h_6	0	0	0	0	0

Figure 2: Tabular representation of a soft set. [8]



Date	MACD_boolean	SO_boolean	RSI_boolean	Sum	Decision
2017-12-07	1.0	1.0	0.0	2.0	Buy
2017-12-08	1.0	1.0	0.0	2.0	Buy
2017-12-11	1.0	1.0	0.0	2.0	Buy
2017-12-12	1.0	0.0	0.0	1.0	Hold
2017-12-13	1.0	0.0	0.0	1.0	Hold
2017-12-14	-1.0	0.0	0.0	-1.0	Hold
2017-12-15	-1.0	1.0	0.0	0.0	Hold

Figure 3: Tabular representation of the soft set used in the model.

provides signals telling brokers to buy while security or currency is oversold and vice-versa to sell while it is being overbought.

$$RSI = 100 - \frac{100}{1 + AverageGain/AverageLoss} \tag{2}$$

where:
$n = 14$ - window length
$AverageGain$ - Sum of gains over the past n periods;
$AverageLoss$ - Sum of loss over the past n periods;

Stochastic Oscillator (SO):

$$\%K = 100\frac{ClosePrice(actual) - MinPrice(n)'}{MaxPrice(n)' - MinPrice(n)'} \tag{3}$$

$$\%D = 100\frac{ClosePrice(actual) - MinPrice(m)'}{MaxPrice(m)' - MinPrice(m)'} \tag{4}$$

$n = 14$ - window length
$m = 3$ - short window length
$ClosePrice$ - Close price from the $actual$ period;
$MaxPrice$ - Maximum price over the past n periods;
$MinPrice$ - Minimum price over the past n periods;

SO functions as indicator of momentum that exploits support and resistance levels. Word stochastic allude to the current price in comparison to its price range over time period. SO tries to foretell price turning points by comparison of security closing price to its price range. Simplifying it looks for the range between an low and high price during time period. Afterwards security's price is expressed as a percentage ranging from 0 to 100. Meaning lower limits of the range over the time period covered, and higher limits over the time period covered. Idea is that prices happen to have tendency to near the extremes of the recent range before turning points. What is actually considered an alert or set-up, is when %D happens to be in an extreme zone and diverging from the price action. Signal itself is happening when the faster %K crosses the %D. Extreme zones are below 20 and above 80.

Moving Average Convergence / Divergence (MACD): Trading indicator developed for technical analysis of stock prices. Gerald Appel created it in the 1970s. It is goal is to describe changes in the direction, strength, momentum, and trend duration in a stock's price. A group of three time series made from past data, most often the closing price. These three are: the proper series , the "signal" series ("average") , and lastly the "divergence" series (difference between the previous ones.) The series is calculated as the difference between a short period exponential moving average, and a longer period exponential moving average of the price series. The average series is an exponential moving average of the MACD series itself. Therefore this indicator stems from three parameters, which are the time constants of the three exponential moving averages. Usually their values are measured in days.

By comparing exponential moving averages of different periods, the moving average convergence / divergence series has the ability to indicate changes in the trend of a stock. As many claim divergence series are a very acute indicator, they can notice subtle shifts in the stock's trend.

Since based on moving averages, it is inherently a lagging indicator. As a metric of price trends, the moving average convergence / divergence is less useful for stocks that are not trending (trading in a range) or are trading with erratic price action.

$$EMA = \frac{p_0 + (1-\alpha)p_1 + (1-\alpha)^2 p_2 \cdots + (1-\alpha)^N p_N}{1 + (1-\alpha) + (1-\alpha)^2 \cdots + (1-\alpha)^N} \tag{5}$$

EMA is the exponential moving average

$\alpha = \frac{2}{N+1}$
p_0 – last value
p_1 – value day-before p_0
p_N value from day N
N – number or periods

$$MACD = EMA(N=12) - EMA(N=26) \tag{6}$$

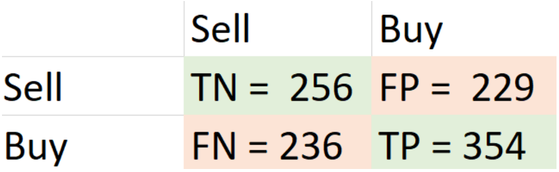
MACD is calculated from the quick and slow EMA difference.

$$singalline = EMA(N=9, values = MACD) \tag{7}$$

Signal-line is created from the calculated on MACD values, with 9 day span.

$$Output = MACD - signalline \tag{8}$$

The actual output is calculated as an difference between MACD and signal-line



	Sell	Buy
Sell	TN = 256	FP = 229
Buy	FN = 236	TP = 354

Figure 4: Error Matrix for evaluation of our method.

IV. Error Matrix

In order to test the model, data is required. NYC stock exchange market provided McDonalds stock prices from the 2013 to 2017. Model provided with data, after few seconds returns decision in an output, in order to check whether his suggestions would make profit, evaluation needs to take place. By checking the stock closing price from the next day, it is possible to see if decision made in day N would make profit (or loss) on day $N+1$. Comparison of output received from the model, and the price from the $N+1$ day takes place in error matrix. With 1075 days total in the data set, model provided 256 profitable sell predictions and 354 profitable buy decisions. Therefore, more then 50% of the suggestions made by model brought profit. In order to understand what happened , why the rest of the data got labeled incorrectly, some statistics need to be calculated. Although, it's better then a coin toss by a small margin, it has capability to become even more accurate. Formulas are shown below (see 9,10,11,12).

$$recall = \frac{TP}{FN + TN} \tag{9}$$

,

$$precision = \frac{TP}{TP + FP} \tag{10}$$

,

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{11}$$

$$f1_{score} = \frac{2TP}{2TP + FP + FN} \tag{12}$$

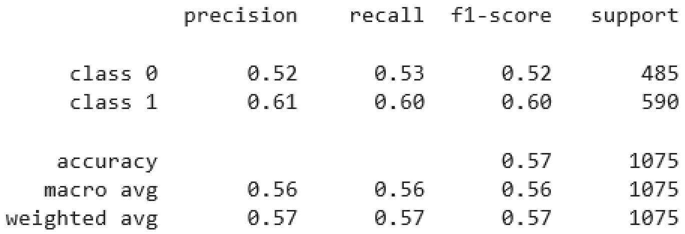
Statistics calculated are shown in figure 5.

As shown in figure 4, scoring 57% (f1 score) at accuracy, means that the model is slightly better then a simple coin toss. Using just three technical indicators, on the stock values of the biggest food company on the world, %57 is a great score. Improving the model with more technical indicators, more sophisticated expert knowledge would allow the model to score even higher.

Simplified flowchart of the model presented in Figure 5 covers the model in the simplest way, without crunching any numbers.

V. Conclusion

Stock market analysis demanding as it is, proven to be possible for models like one presented in this issue. Soft sets are a great choice if expert knowledge and broad choice of parameters is available. That being said, some facts must be taken into consideration. Plenty of different elements put



```
              precision    recall  f1-score   support

     class 0       0.52      0.53      0.52       485
     class 1       0.61      0.60      0.60       590

    accuracy                           0.57      1075
   macro avg       0.56      0.56      0.56      1075
weighted avg       0.57      0.57      0.57      1075
```

Figure 5: Statistics of the matrix results

pressure on the stock market. Working just on the price values, is a very narrow way of analysis. Although at the first glance it is all about numbers, there is more to it then covered in this model.

References

[1] F. Beritelli, G. Capizzi, G. L. Sciuto, C. Napoli, and M. Woźniak, "A novel training method to preserve generalization of rbpnn classifiers applied to ecg signals diagnosis," *Neural Networks*, vol. 108, pp. 331–338, 2018.

[2] M. Woźniak and D. Połap, "Hybrid neuro-heuristic methodology for simulation and control of dynamic systems over time interval," *Neural Networks*, vol. 93, pp. 45–56, 2017.

[3] D. Molodtsov, "Soft set theory—first results," *Computers & Mathematics with Applications*, vol. 37, no. 4-5, pp. 19–31, 1999.

[4] H. Aktaş and N. Çağman, "Soft sets and soft groups," *Information sciences*, vol. 177, no. 13, pp. 2726–2735, 2007.

[5] F. Feng, X. Liu, V. Leoreanu-Fotea, and Y. B. Jun, "Soft sets and soft rough sets," *Information Sciences*, vol. 181, no. 6, pp. 1125–1137, 2011.

[6] F. Fatimah, D. Rosadi, R. F. Hakim, and J. C. R. Alcantud, "Probabilistic soft sets and dual probabilistic soft sets in decision-making," *Neural Computing and Applications*, vol. 31, no. 1, pp. 397–407, 2019.

[7] J. Hu, L. Pan, Y. Yang, and H. Chen, "A group medical diagnosis model based on intuitionistic fuzzy soft sets," *Applied Soft Computing*, vol. 77, pp. 453–466, 2019.

[8] P. K. Maji, R. Biswas, and A. Roy, "Soft set theory," *Computers & Mathematics with Applications*, vol. 45, no. 4-5, pp. 555–562, 2003.

[9] D. Pei and D. Miao, "From soft sets to information systems," in *2005 IEEE international conference on granular computing*, vol. 2. IEEE, 2005, pp. 617–621.