**Capstone Project** Theo Lekkas

Machine Learning Engineer Nanodegree July 2016

**Definition**

**Project Overview**

In this project I look to determine whether or not the Relative Strength Index (RSI) is an effective tool for buying & selling stocks. First I begin with a definition of the RSI:

“The relative strength index (RSI) is a technical momentum indicator that compares the magnitude of recent gains to recent losses in an attempt to determine overbought and oversold conditions of an asset.”[[1]](#footnote-1)

The RSI is supposed to be able to determine both good buy & sell points for stocks (it is also used in trading futures, but I will restrict my analysis to stocks for this project.) The purpose of this project is to determine whether or not buying a stock when it’s oversold & then selling the stock when it’s overbought is a profitable strategy. Below, I will more clearly state the parameters of this problem & further clarify various definitions.

**Problem Statement**

This project intends to determine whether a trading strategy of going long (buying) a stock when it is considered oversold using the RSI is a viable (profitable) strategy. Profitability, however, is not enough for a viable strategy. We will also want a comparison against some benchmark; the reason for this is to determine whether or not it was better to simply invest in the broader market rather than deploying a specific trading/investing strategy. For this project I will use the returns to the S&P 500 over the time period used for this strategy. There are other more sophisticated metrics that look at the volatility of the strategy (such as Sharpe ratios), but for two reasons I will not be looking at these: 1) for the purposes of this project I am only interested whether or not RSI is really predictive and nominally better than the broader market 2) I don’t much care about volatility, to quote Warren Buffet “I would much rather earn a lumpy 15% over time than a smooth 12%.”[[2]](#footnote-2)

The buy signal for a stock will be when the RSI indicates an oversold threshold has been reached & the sell signal will be when the RSI indicates an overbought threshold has been reached. I will use adjusted daily prices. Both the purchases & sales will be made on the adjusted closing price. In the initial attempt I will not consider commissions, slippage, or other market frictions; these can be added later if the strategy actually seems viable. Furthermore, I will also look at various moving averages, which the RSI is dependent upon. I will also look at different levels of oversold & overbought (these will become clearer below when the RSI formula is broken down.)

I will use two different machine learning models, linear regression & decision trees, to see if there is predictive value to this strategy.

Below I will more clearly define RSI and various elements/concepts that will be used throughout the project.

**Metrics**

The RSI is a momentum indicator that attempts to determine short-term bullish (positive) & bearish (negative) positions for stocks. The RSI looks at the price action of a stock and does not consider other elements (such as sentiment, news, accounting metrics, etc.) The RSI is a pure “technical” indicator, its only concerned with the price movement of the instrument being analyzed.

The formula for the RSI is as follows:



Created with <http://www.sciweavers.org/free-online-latex-equation-editor>

As can be interpreted from the formula above, the RSI is an index between 0 & 100.

The RS is the relative strength metric, it is defined as average of up days over n periods divided by the average of down days over n periods. There are a couple of ways to calculate the averages; smoothed moving average[[3]](#footnote-3) or exponential moving average[[4]](#footnote-4). Below is formula for the RS using a smoothed average:



Created with <http://www.sciweavers.org/free-online-latex-equation-editor>

The U & D in the above formula represent ‘Up’ & ‘Down’ which are calculated by subtracting the current close with the previous days close[[5]](#footnote-5).

Finally, we need to define the meaning of Oversold & Overbought. These are somewhat arbitrary, but for the initial runs I will use the traditional definitions. Oversold on the RSI is < 30 and Overbought > 70. These two definitions are definitely ripe for fine-tuning, but we will use the traditional definitions at first.

**Analysis**

**Data Exploration**

The data set used is taken from the Quandl WIKI data[[6]](#footnote-6). This is a community curated set of 3179 stocks (as of 06/08/2016, the date the dataset used in this project was downloaded.) The data set consists of the entire[[7]](#footnote-7) price history of the stocks included in the WIKI data set. The price history includes the open, high, low, & close as well as adjusted prices (adjusted prices take into account stock splits & dividends and update the historical prices in order to have continuous prices where there are no large changes to prices that had nothing to do with daily changes.) The data also includes the ticker symbol, the date, and the volume (as well as adjusted volume.) The dataset itself is freely available from the Quandl website (linked to in the footnotes), but requires an API key (which is free, but requires registration.) The WIKI dataset, as of download date, is a csv file with 14,150,092 rows of data with 14 columns to each row.

The subset of data used consists of the adjusted price/volume data, the ticker symbol, and the date. Furthermore, there will be a subset of the entire WIKI dataset included in the repository. The entire WIKI dataset is ~1.6 gb in size. The subset will include the stocks that were used in the Jupyter notebook that accompanies this report as well as well number of other stocks randomly chosen in case some additional analysis/testing is desired on the part of the reader.

All stock datasets are in chronological order and then organized alphabetically. The stocks used are taken from the larger dataset using a ticker\_gettr() function that is defined in the notebook.

**Exploratory Visualization**

**Algorithms and Techniques**

**Benchmark**

As with most trading strategies the benchmark to compare the strategy to is usually the S&P 500 returns. In this instance, not only do I want to compare the strategy of buying a stock below RSI 30 & closing the position (i.e. selling the stock) above RSI 70 to the S&P 500, but I also want to compare this strategy to simply going long the stock over the time frame the strategy was being executed. By doing so I can make two necessary comparisons. The first comparison is whether or not the strategy had better returns than simply going long the individual stock (“buy and hold”.) This is an important comparison to make since the buy & hold strategy has significantly lower transactions/slippage costs than any trading strategy does. So if a trading strategy can perform better than simply buying & holding the security trading, then the strategy has some viability (since the returns would be higher.)

The second comparison, to the S&P 500[[8]](#footnote-8) (the Wilshire 5000[[9]](#footnote-9) index is sometimes used for this type of comparison as well as the index represents the all publicly traded U.S. equities), is an important comparison because buying the S&P 500 represents going long the U.S. markets and diversifying away any industry or company specific risk. The argument for using the S&P 500 as a benchmark is that if a strategy can’t return more than a simplistic, low risk buy & hold strategy using the S&P, then the trading/investing strategy is not worth the additional costs (i.e., management fees, commissions, slippage, etc.) associated with an active strategy.

Holding periods.

**Methodology**

**Data Preprocessing**

A number of preprocessing steps were taken in order to build a dataset that would be used to build the machine learning models. The preprocessing steps are clearly shown in the accompanying notebook, but will be discussed here in order to make certain that reasoning behind each step is clear.

The first step was to create a new dataframe from the dataset that was opened initially. The purpose of this step was to make use of all the adjusted data. The reason I chose the adjusted data is relatively straightforward. The adjusted data adjusts for changes in price that are due to splits or dividends. Therefore, adjusted datasets for stock prices are fairly continuous, in that these datasets only show prices fluctuations that are due to market changes.

The next step was to add a column for the RSI data that is calculated in the notebook. This step also requires the removal of some the earliest rows depending on what moving average is being used (e.g. the first 14 rows are removed if the 14 day moving average is used.)

The next step adds two new columns. These columns are the ‘Buy Price’ & ‘Sell Price’ columns. The purpose of these columns is to get the adjusted closing price of the stock when the stock is less than the low threshold (the ‘Buy Price’) or greater than the high threshold (the ‘Sell Price’) for the RSI.

The next step was to add a number of columns that represent holding periods. Rather, these columns have the adjusted closing price a number of days in the future. The purpose of these columns was to see if there was a simple holding period strategy that was/is better than using the upper RSI threshold.

The final step was to add a column for the ‘Sell Signal Price’. This column represents the adjusted closing price for the first > RSI 70 day after the initial buy signal. This is place to discuss what is going on in this column. There are many instances of the same price being used in the in the ‘Sell Signal Price’ column. The reason why there is often the same price is used is because I am looking for the first instance *after* of sell signal the buy signal is generated. Often times there will be multiple instances of the buy signal being generated before a sell signal being generated. All of the buy signals are still valid, they just happen to have the same sell signal.

**Implementation**

**Refinement**

**Results**

**Model Evaluation and Validation**

**Justification**

**Conclusion**

**Free-Form Visualization**

**Reflection**

**Improvement**

1. Relative Strength Index - RSI <http://www.investopedia.com/terms/r/rsi.asp> [↑](#footnote-ref-1)
2. 1996 Chairman's Letter - <http://www.berkshirehathaway.com/letters/1996.html> [↑](#footnote-ref-2)
3. [https://en.wikipedia.org/wiki/Moving\_average - Simple\_moving\_average](https://en.wikipedia.org/wiki/Moving_average#Simple_moving_average) [↑](#footnote-ref-3)
4. [https://en.wikipedia.org/wiki/Moving\_average - Exponential\_moving\_average](https://en.wikipedia.org/wiki/Moving_average#Exponential_moving_average) [↑](#footnote-ref-4)
5. [https://en.wikipedia.org/wiki/Relative\_strength\_index - Calculation](https://en.wikipedia.org/wiki/Relative_strength_index#Calculation) [↑](#footnote-ref-5)
6. Quandl WIKI data: <https://www.quandl.com/data/WIKI> [↑](#footnote-ref-6)
7. For certain stocks not the entire stock history is available in this dataset. [↑](#footnote-ref-7)
8. It is important to note that this is U.S. biased research, but the principles of the strategy can be translated to any financial market. [↑](#footnote-ref-8)
9. <https://en.wikipedia.org/wiki/Wilshire_5000> [↑](#footnote-ref-9)