Shane Bari Russ Chamberlain Cassie Dale Jared Siverling

STATUS REPORT 1

10/8/15

INTRODUCTION

Client Objective:

The following project has been motivated by Investment Counsel LLC. Investment Counsel LLC requests a computer program that would work as an investment tool and predict whether to buy or sell stocks and help Wealth Management workers better allocate funds to maximize profit and minimize loss. The program should be automated and require no supervision, be able to adjust risk tolerances, and have automated purchase and sell triggers that are set by percentages scaled by relative stock or fund values.

Problem Description:

The stock market is large, unstable, and unpredictable, as such, a human cannot efficiently or effectively analyze all significant data available to determine a stock's potential value and its potential variation. Therefore, a program that takes in large amounts of data and predicts when to buy/sell/hold and alerts investors of risk is essential to optimize profit and minimize loss. The goal of this project is to research potential triggers of an individual stock(s), develop a program that meets client's specifications, and investigate difficulties, limitations, and concerns of such a program.

Considerations:

Efficient Market Hypothesis:

Efficient Market Hypothesis states that the future stock price is completely unpredictable given the past trading history of the stock. There are three perspectives to this hypothesis: strong, semi-strong, and weak. The strong EMH contends that stock prices do not depend on past stock prices, so patterns cannot be exploited since trends do not exist. On the other hand, the weak EMH says that any information acquired from examining the stock's history can be reflected in the price of the stock.

Fundamental/Technical Analysis:

Fundamental Analysis: this type of analysis is concerned more with the company than the actual stock. Decisions are made based on the past performance of the company, the earning forecast, etc.

Technical Analysis: this method deals with the determination of the stock price based on the past patterns of the stock, using time-series analysis.

When applying Machine Learning to Stock Data, we are more interested in doing a Technical Analysis to see if our algorithm can accurately learn the underlying patterns in the stock time series. This said, Machine Learning can also play a major role in evaluating and forecasting the performance of the company and other similar parameters helpful in Fundamental Analysis. In fact, the most successful automated stock prediction and recommendation systems use some sort of a hybrid analysis model involving both Fundamental and Technical Analysis.

Model Details:

Indicators:

Several indicators can be used in the analysis of stock prices, these indicators will be used to set triggers:

Moving Average: the average of the past *n* values until today

Exponential Moving Average: gives more weightage to the most recent values while not discarding the older observations entirely

Rate of Change: the ratio of the current price to the price *n* quotes earlier

Relative Strength Index: Measures the relative size of recent upward trends against the size of downward trends within a specified time interval

Machine Learning:

Several studies have utilized Machine Learning techniques to predict the rise and fall of stock prices before the actual increase or decrease in stock prices occurs. The following algorithms have proven to be moderately successful at making accurate predictions, such as:

- Artificial Neural Networks
- Support Vector Machines
- Regression (linear, logistic, etc.)
- K-Nearest Neighbors
- Decision Trees

The exact combination of algorithms will be determined by processing historical data and selecting important attributes. The processed data is then split into two subsets: the *training data* and *test data.* The training data is fed through the algorithm in an iterative process as it is "learned" by the ML program the learning has completed, *Cross Validation* can be used to evaluate the performance of the model versus the left over test data. If the model is not satisfactory given the test data, a new algorithm can be trained; otherwise the model can now be used for prediction.

Tools

Software:

The following software will be utilized during this project:

- R (Statistical Programming, load data, plots)
- Github (Share and Store code)
- SQL (Store data)
- Python (More programming, user interface)
- Mircosoft Excel (Grab data, plots)

R will be used for the required statistics, probability, and computations. Github will be used to store code and files. SQL will be used to manage, transfer, sort, and edit data. Python will be used to attempt to set up a more user friendly program, user interface, instead of just an algorithm in R, if a successful algorithm is found. R and Microsoft Excel will also be used when plotting data.

The following software may be utilized as needed:

- Sage/Maxima (computations)
- SPSS (Statistical programming)

Data:

Technical data will be pulled primarily from Yahoo Finance. A text base search for triggers may also be incorporated that would involve most of the web.

Data of vocabulary is necessary for an understanding and background of investing. To learn necessary background information we will be using the following:

- Merrill Edge account
- My Nastaq
- General Web search
- Investopedia

Math/Stats/Methods:

The project will require various tools from statistics and mathematics. Statistics will be used to access significance, test hypothesis, identify trends, and make predictions, confidence intervals. Probability will be utilized in predicting risk and appropriately modifying the tolerances of the algorithm. A listing of the various methods and subjects to be used follows:

- Linear Regression
- Moving averages
- Time series
- Seasonality
- Probability
- Checking error
- Testing significance
- Smoothing
- Data reduction

Initial Model and Considered Solutions

Ideas:

There are two main ideas for how to approach this problem: predicting with past data and predicting with current data. When looking for triggers, we can focus on past data and try to find trends between/among similar companies and make inferences on how stock will fluctuate based on precedents from companies that have already gone through similar changes. That is, if we can show that a current company mimics or is trending like a company 5 years ago and is about to make similar changes or take similar actions, we can use the past companies data to make predictions on the current company's outcome, using the precedent of the first company it is mimicking. The other way we can look at triggers is testing the significance of current events on the market value of the stock, this may be harder to isolate and prove though.

Initial Models:

(1) An algorithm in R that takes in data from Yahoo Finance, considers rolling averages to make predictions, run error analysis and check significance, and modifies itself from past errors.

(2) A algorithm in R that looks for trends among companies with past data, categorizes companies based on trends and similarities, makes predictions from precedents of similar past companies, and adjusts based on error.

(3) A algorithm in R that integrates the first two models and utilizes triggers based on precedents of similar past companies and current events. This would most likely be the most accurate model but would require the most work in merging the two different types of triggers and predictions to make sure they do not interfere with one another.

Limits of Initial Models:

The initial models will be primarily concerned with daily rates, this may be expanded to weekly but any larger amount of time will not be predicted by the algorithm. The algorithm will, initially, focus on Large Cap Large Value companies that are less chaotic and have a plethora of available data. Because of this, the algorithm will not be effective for smaller or more random companies. The models will also not consider catastrophic events such as but not limited to market shut down, hacked news networks, states of annihilation, etc. More limits would be considering the assumptions of statistical models.

Tolerances:

Allowing a tolerance input will involve an interface for the user or some basic R training to allow the user to modify or define their desired tolerance to risk. The model must allow for more or less risk depending on the user's input. This will be achieved by focusing on predicting potential error and utilizing the bounds of confidence intervals.