Lamberto Pisani

Proposal

Domain Background:

Today's called Efficient Market Hypothesis (EMH) based on the research of Samuelson (1965) and Fama (1965) states that Stock markets behavior and more importantly, stock prices behavior, follow a "random walk" if investors behaved rationally (Shleifer and Summers, 1990); Shleifer And Summers (1990) stated that this theory lost grounds rapidly thanks to publications such as Shiller's (1981), Leory and Porter's (1981) and it particularly crashed after the market crash on October 19, 1987.

In other words, according to the EMH (Nardo, Petracco-Giudici and Naltsidis, 2016) financial prices are not predictable, however, many attempts have been made to predict them. More recently, thanks to technological and data processing advances new machine learning(ML) techniques have been used because these allow to uncover generalizable patterns not specified in advance (Mullainathan and Spiess, 2017). Some examples are: directional prediction (classification) based on news contents (Alsotad and Davalcu, 2017) and stock daily returns (Gunduz, Cataltepe and Yaslan, 2017 and Li, Xie and Wang, 2016) to mention a few.

This project intends to apply ML classification techniques to predict the directional movement of a selected company's stock or ETF. From a personal point of view, this study and its outcome(s) will help the author of this work to start a personal investment strategy based on these initial results and further improvements over time; this is because current investment alternatives (financial advisors, etc) offer high prices/costs for their services, however, they don't share the risks of the investment strategy suggested to their clients (a.k.a. the low side).

Performing this project will also train the author to work with time series data as well as understanding basic concepts of stock trading such as data preparation and processing.

Problem Statement:

This project's objective is to define directionality of the closing price of a stock/ETF *N* days in the future relative to its daily opening price i.e. whether its price is predicted to increase (up), or decrease (down). Given the binary nature of the outcome, supervised learning techniques will be used to achieve the objective; specifically, classification techniques such as Decision Tress, Gaussian Naïve Bayes, Support Vector Machine and Deep Networks (in order of complexity) will be used.

Datasets and Inputs (General View):

Basic stock market information is readily available for consumers in websites such as https://finance.yahoo.com/, www.google.com/finance or https://www.quandl.com/ to mention a few. There are also curated data sets in sites such as Kaggle.com. In particular, this project will utilize the Huge Stock Dataset as it already includes not only one stock, but also ETFs data as far back as 1999 depending on the selected stock. In particular, for this analysis the following inputs and/or data sets will be used:

Moreover, for this project the following datasets will be used as part of the analysis for each chosen stock/ETF and Index:

- "Daily and Intraday Stock Price Data" from Kaggle.com (https://www.kaggle.com/borismarjanovic/daily-and-intraday-stock-price-data). It contains full historical daily (an intraday) price and volume data for all U.S.-based stocks and ETFs trading on the NYSE, NASDAQ, and NYSE MKT (Date, Open, High, Low, Close, Volume, OpenInt). This data set is adjusted for splits and dividends. Specifically, the daily data will be used.
- NASDAQ-100 Technology Index (NDXT) https://finance.yahoo.com/quote/%5ENDXT/history?p=^NDXT
- Volatility Index (VIX) http://www.cboe.com/products/vix-index-volatility/vix-options-and-futures
    - 2004 through Jan 19, 2018
- NASDAQ Volatility Index (VXN) http://www.cboe.com/products/vix-index-volatility/volatility-on-stock-indexes/cboe-nasdaq-100-volatility-index-vxn
    - 2001 through Jan 19, 2018
- Engineered Features (as per definition on Madge, 2015):
    - Up_Down: defines whether the stock increased or decreased its price each day
    - Security's Price Momentum
    - Security's Price Volatility
    - Sector's Momentum - NDXT
    - Sector's Volatility - NDXT
    - Market's Momentum - VIX
    - Market's Volatility - VIX

These datasets will be combined into one dataframe for easy training and testing afterwards. This is, data from VIX, NXDT and VXN data sets will be brought in into the Stock Data set, then time series will be detrended, features scaled and finally, engineered features will be calculated. Once combined, the data will be separated into Training, Validation and Testing dataset for the training and evaluation process.

Datasets and Inputs (Detailed View):

The author of this project is not experienced on investing techniques, hence, a "safer" approach will be used, this is, this project will be based in the selection of ETFs from a set of 49 technology sector ETFs (author's background) based on their current performance given on the following link (on 1/25/18): 49 Best ETFs. From these 49 ETFs a subset of 8 was selected so the following are met:

- ETFs are available in "Huge dataset"
- Data available for ALL ETFs and market indexes as far back as possible (2/23/2006)

The final combined dataset contains the following engineered features:

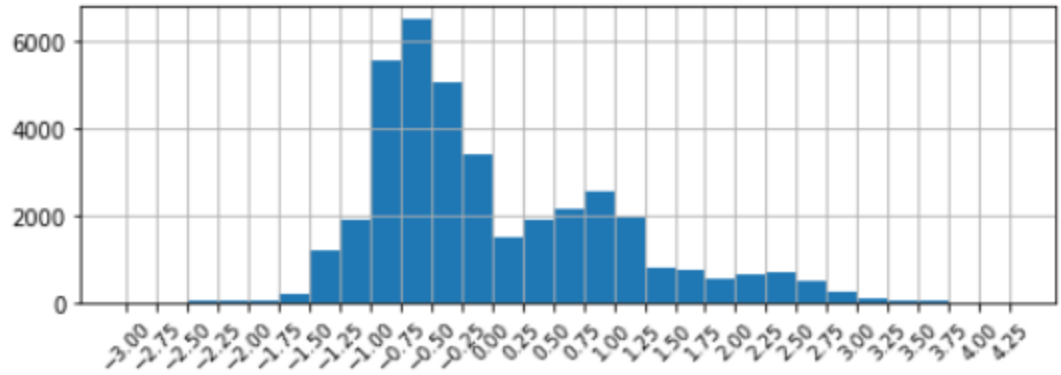
- 13 of the 49 ETFS

- Daily difference calculation for each Price value Colum as well as market index values. These differences (a.k.a delta values) considered the grouping of each “Ticker”(ETF) so as to avoid creating deltas coming from values in other groups
- Standardized values for each “Price”/Index value as well as for the Delta calculated columns. These will allow to understand what to expect in terms of the balance of the features

After a brief look at the results of the dataset:

1) The dataset contains 38921 rows for 13 “Tickers” with a data range from 2/23/2016 through 11/10/17.
2) Delta Price values (target feature) for all 13 selected “tickers” have a tendency to the “low” side with a total average of around 62.48% of the sample.

```
Index to be Plotted CLOSE
-2.5    -   4.0
```



When looking at each individual ticker, the tendency stays with only two tickers showing down trends lower than 55%.

Ticker	Down	Up	Grand Total
IGM	62.35%	37.65%	100.00%
IGN	54.52%	45.48%	100.00%
IGV	62.15%	37.85%	100.00%
IXN	64.28%	35.72%	100.00%
IYW	63.57%	36.43%	100.00%
PSI	67.78%	32.22%	100.00%
PSJ	60.43%	39.57%	100.00%
PXQ	52.48%	47.52%	100.00%
SMH	65.10%	34.90%	100.00%
SOXX	66.57%	33.43%	100.00%
VGT	63.03%	36.97%	100.00%
XLK	62.28%	37.72%	100.00%
XSD	67.53%	32.47%	100.00%

Grand Total	**62.48%**	**37.52%**	**100.00%**

- Because of the results above and considering that ETFs are not day trading instruments but they have a longer investment horizon, a new category has been added to account for Up/Down and No changes where the "No Change" threshold was selected as centered around the mean +/- 0.5*Std of each sample/row (the standardized feature was calculated as part of the feature engineering explained above and visualized in Excel for simple manipulation). Using this approach, the following summary is obtained.

Ticker	Down	No Change	Up	Grand Total
IGM	43.99%	25.86%	30.16%	100.00%
IGN	31.65%	41.04%	27.31%	100.00%
IGV	42.19%	28.30%	29.52%	100.00%
IXN	41.61%	28.91%	29.48%	100.00%
IYW	42.26%	27.99%	29.75%	100.00%
PSI	35.62%	41.14%	23.24%	100.00%
PSJ	42.04%	27.28%	30.68%	100.00%
PXQ	39.16%	28.57%	32.26%	100.00%
SMH	40.77%	31.55%	27.69%	100.00%
SOXX	37.03%	36.69%	26.29%	100.00%
VGT	43.58%	26.47%	29.96%	100.00%
XLK	45.34%	24.97%	29.68%	100.00%
XSD	39.53%	33.93%	26.54%	100.00%
Grand Total	**40.37%**	**30.98%**	**28.65%**	**100.00%**

The results above show an imbalance either as a whole group or by individual ticker names. Looking at them, the tickers with the "highest" UP probabilities are: PXQ, PSJ and IGM with over 30% of the samples above MEAN + 0.5*(Std of the group)

Solution Statement

A directional prediction will be developed for N days in the future given the data above. Supervised Classification techniques will be used using as label the UP_DOWN calculated feature. Multiple Algorithms will be evaluated for this project (Decision Tress, Gaussian Naïve Bayes, Support Vector Machine and Deep Networks) using standard evaluation metrics.

Lamberto Pisani

Benchmark Model

Shleifer and Summers (1990) defined stock predictions as a "random walk". This basically means that any classification prediction made will have an expected average of 50% of success, however, from the balance results obtained above this is not the case, hence, a purely naïve approach may not be providing the best results. Therefore, a machine learning approach will be used, this is, the benchmark model will be defined as an out-of-the-box Random Forest on the project data. By using this approach, a like-for-like comparison can be made between the Random Forest and the trained model of this project.

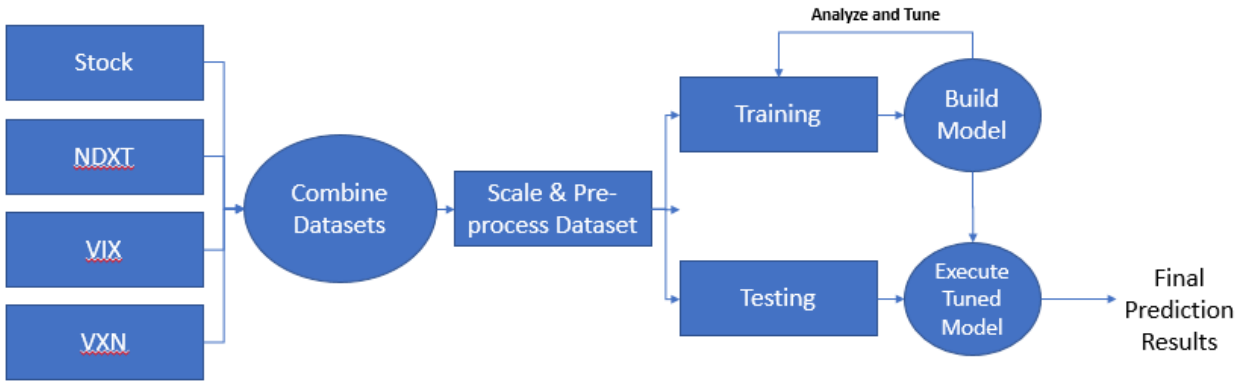
Evaluation Metrics

It was first intended to use cross-validation for the model's evaluation and selection; however, since this project deals with time series, techniques such as standard k-fold cross validation may not provide the best results given the intrinsic nature of time dependency on time series variables which is missed by the standard procedure (https://stats.stackexchange.com/questions/14099/using-k-fold-cross-validation-for-time-series-model-selection, https://blog.insightdatascience.com/whats-wrong-with-my-time-series-model-validation-without-a-hold-out-set-94151d38cf5b). Because of this, the author will use a cross-validation technique specific for time series in the Sklearn library (http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.TimeSeriesSplit.html)

Given the classification requirement of this project and its domain, historically many evaluation metrics have been used for measuring prediction's success including accuracy and F-Scores. More than half of the previous research studies used accuracy as evaluation metric, however, accuracy has predictive power on only one (the positive) of the possible choices (Fortuny et all, 2014). Another evaluation metric used is the Area Under the Curve of the Receiving Operating Curve (ROC) which allows the evaluation of a model when the target labels have a skewed distribution (Fortuny et all, 2014). Other metrics can also be used such as trading simulation and Sharpe ratio, but this project will focus on ROC.

Project Design

Below a high-level view of the process to follow for the realization of this project.



Below a brief explanation for this process:

1) The raw datasets will be read and combined into one dataset
2) The data will be pre-processed for assessing balance:
    a. Timeframe will be based on the maximum timeframe available for all datasets
    b. Calculate Engineered features:
        i. "Closing" ticker prices will be used as target feature for defining daily directionality
        ii. Delta daily prices for each "ticker"
        iii. Standardize values and evaluate balance as well as define Up, Down and No changes. "No change" is used to avoid too many transactions due to the long-term investment horizon of ETFs
3) New set of features will be calculated:
    a. Price Volatility for each Market Index Used
        i. Based on different timeframe segments for later evaluation
    b. Price Momentum for each Market Index Used
        i. Based on different timeframe segments for later evaluation
    c. Sector Volatility for each Market Index Used
        i. Based on different timeframe segments for later evaluation
    d. Sector Momentum for each Market Index Used
        i. Based on different timeframe segments for later evaluation
4) Scaling all features to keep same scale for training purposes
5) Split the data to use Cross-Validation for time series
6) Training Model for each lag timeframe combination
7) Tuning the Model using classifiers Decision Tress, Gaussian Naïve Bayes, Support Vector Machine and Deep Networks for each lag timeframe combination
    a. Select the model/lag timeframe that provides best prediction results
    b. Select the Market Indexes that provide better prediction results
8) Test model using test dataset
9) Compare the model results with benchmark (50% success)

As part of the preparation process, it is considered to use feature reduction techniques (Such as PCA) to quantify or assess which features provide better prediction results.

References:

- Alsotad and Davalcu (2017). Directional prediction of stock prices using breaking news on Twitter. Web Intelligence and Intelligent Agent Technology (WI-IAT), 2015 IEEE / WIC / ACM International Conference. http://ieeexplore.ieee.org/document/7396858/?reload=true
- Fama, E. (1965). The Behavior of Stock-Market Prices. Journal of Business, Volume 38, Issue 1 (Jan., 1965), 34-105.
- https://www.jstor.org/stable/2350752?seq=1#page_scan_tab_contents
- Fortuny et all (2014). Evaluating and understanding text-based stock price prediction models. Information Processing and Management 50 (2014) 426–441. http://www.sciencedirect.com/science/article/pii/S0306457313001143?via%3Dihub
- Gunduz, Cataltepe and Yaslan (2017). Stock daily return prediction using expanded features and feature selection. Turkish Journal of Electrical Engineering & Computer Sciences (2017) 25: 4829 – 4840. https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=0ahUKEwjo7NmyxurYAhVR7qwKHfpYAyEQFgguMAA&url=http%3A%2F%2Fjournals.tubitak.gov.tr%2Felektrik%2Fissues%2Felk-17-25-6%2Felk-25-6-32-1704-256.pdf&usg=AOvVaw3Rw6mCh3x1rWkD2fGjHaWZ
- Samuelson, Paul A., Proof That Properly Anticipated Prices Fluctuate Randomly Industrial
- Management Review, 6:2 (1965:Spring). http://jrjgb.jj.cqut.edu.cn/__local/6/35/CE/516B5F529EC5AAF4B9C1FFD5C4B_A532FC8A_B39E2.pdf
- Shleifer and Summers (1990). The Noise Trader Approach to Finance. The Journal of Economic Perspectives, Vol. 4, No. 2 (Spring, 1990), pp. 19-33. http://www.jstor.org/stable/1942888
- Madge, Saahil (2015). Predicting Stock Price Direction using Support Vector Machines
- https://www.cs.princeton.edu/sites/default/files/uploads/saahil_madge.pdf
- Mullainathan and Spiess (2017). Machine Learning: An Applied Econometric Approach. Journal of Economic Perspectives—Volume 31, Number 2—Spring 2017—Pages 87–106. https://www.aeaweb.org/articles?id=10.1257/jep.31.2.87
- Nardo, Petracco-Giudici and Naltsidis (2016). Walking down wall street with a tablet: a survey of stock market. Journal of Economic Surveys (2016) Vol. 30, No. 2, pp. 356–369. http://onlinelibrary.wiley.com/wol1/doi/10.1111/joes.12102/full
- Li, Xie and Wang (2016). Empirical analysis: stock market prediction via extreme learning Machine. Neural Comput & Applic (2016) 27:67–78. https://link.springer.com/article/10.1007/s00521-014-1550-z