A constituent sentiment approach to stock market trend prediction

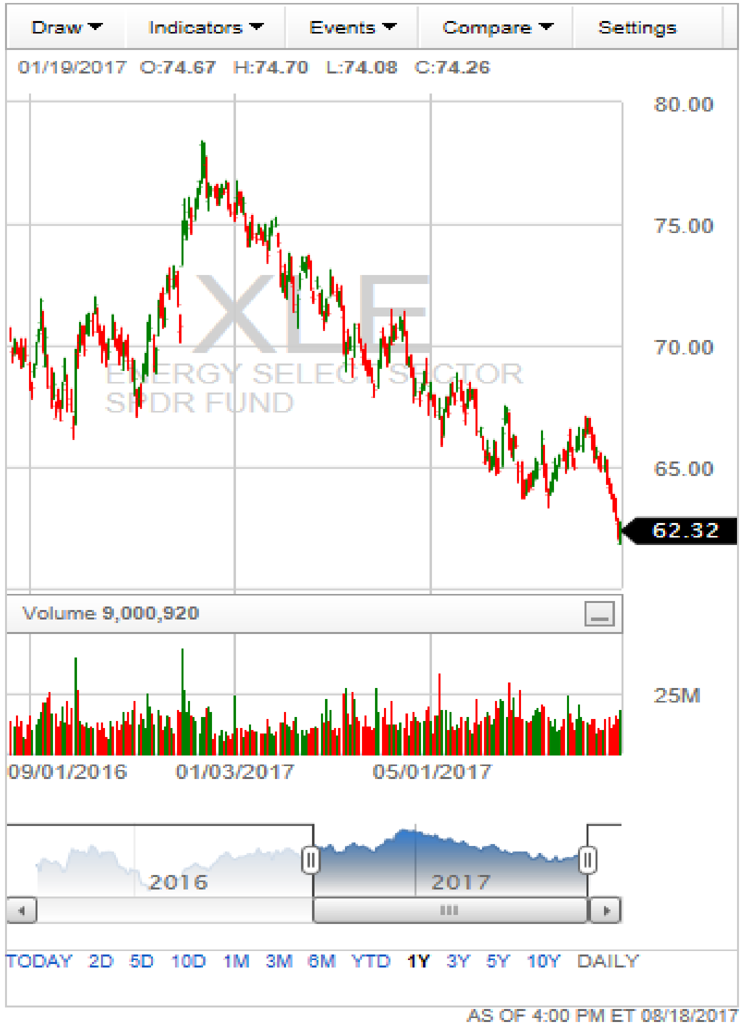
by Frédéric Maréchal

September 2017

Thesis Supervisor: Dr Daniel Stamate

Definition

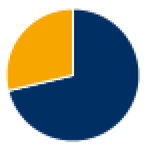
Stock market index is computed from the weighted average price of its constituents.



Source: https://screener.fidelity.com/ftgw/etf/goto/snapshot/snapshot.jhtml?symbols=XLE

Basket Holdings

Total: 33 (Long: 32 | Short: 0)



Top 10	71.27%
Other	28.73%

Top 10

XOM	Issuer Name	22.37%
CVX	Issuer Name	16.96%
SLB	Issuer Name	7.48%
COP	Issuer Name	4.59%
EOG	Issuer Name	4.19%
OXY	Issuer Name	3.92%
PSX	Issuer Name	3.13%
KMI	Issuer Name	3.11%
HAL	Issuer Name	2.96%
VLO	Issuer Name	2.56%

View all Basket Holdings by Weight

Source: https://screener.fidelity.com/ftgw/etf/goto/snapshot/portfolioComposition.jhtml?symbols=XLE

Project Scope

- Does the XLE constituents’ sentiment has predictive power over the XLE index?

- Does the XLE sentiment has a predictive power on the XLE trend prediction?

- Can the sentiment improve the Index volatility prediction?

Contribution to knowledge

- We propose a new framework to integrate sentiment to trend/volatility prediction

- It can be used as part of index arbitrage or portfolio allocation strategies as part of a fully automated trading tool or an extra advising tool.

- Better volatility prediction helps with risk management and reducing capital allocation.

Literature Review

Perfect information & the Random Walk

- The efficient market hypothesis theory (EMH) (Fama,1969)
- The random walk hypothesis, (Malkiel,1973)

Existence of Market anomalies?

- Volatility spikes
- Sudden and sharp regime change (e.g. crisis)
- The practical evidence of cross-sectional pricing anomalies (Keim,2006)

The common market trend explanatory variable types

- Fundamental analysis indicators
- Technical analysis indicators

Literature Review (ctn’d)

ML & Technical/Fundamental Indicators	
Technical analysis indicators predicators	**Vaiz and Ramaswami (2014)** - Predicators: 20 technical indicators (e.g. : RSI, EMA, MCDA, etc.) - Models: Decision Tree, CART and C5.0 with a single training/test sets - Results: Avg 85% accuracy in predicting mkt trend **Parikh and Shah P (2015) & Senyurt and Subasi (n.d.)** - Predicators: Technical indicators - Models (results): Decision Tree (Avg 80%), Random Forest (Avg 79%), Naive Bayesian classifiers (Avg 74%) with 10-fold cross-validation
Fundamental analysis indicators predicators	**Joshi et Al (2013) & Imandoust and Bolandraftar (2014)** - Predicators: Fundamental indicators - Models: Random Forest with single training/test sets - Results: Avg 62% accuracy in predicting mkt trend

Literature Review (ctn'd)

ML & Sentiment	
Sentiment predicators	**Meesad and Li (2014) & Schumaker and Chen (2009)** - Bag-of-words approach from tweets - Feature selection - Corpus to extract sentiment score - Generation of sentiment weights used as attributes $\mathrm{W}ij = \begin{cases} vij + senti(ti), & senti(ti) > 0 \\ -1 * vij + senti(ti), & senti(ti) < 0 \end{cases}$ where senti(ti) = (∑ score(pos) - ∑ score(neg)) / n vij: tokens weight, defined by the Term Frequency-Inverse Document Frequency (TF-IDF) The response variable is the $Trend = \begin{cases} up, & price\ today - price\ yesterday > 0 \\ down, & price\ today - price\ yesterday < 0 \end{cases}$ - Model: SVM with a single training/test set for Parikhs and Shah P (2015) and Leave-One-Out cross-validation for Schumaker and Chen (2009) - Result: 93.4% of accuracy rate (trend prediction)

Literature Review (ctn'd)

ML & Technical/Sentiment Indicators	
Technical and Sentiment predicators	**Halgamuge (2007)** - Methodology: Bag-of-words generated from news article + technical indictors - Model: SVM with a single training/test sets. - Results: 58.8% test accuracy (technical indicators only) 62.5% test accuracy (company news only) 64.77% test accuracy (the company and market news) 70.1% test accuracy (the price, the company and market news)

Literature Review (ctn'd)

A Statistical Approach

- Olaniyan, Stamate and Logofatu (2015) expanded on the previous research from Gilbert and Karahalios (2010) classified *20 millions posts from Livejournal* on the S&P500. They used a linear Granger causality test on two Vector Autoregression models (M1/M2), to prove that anxiety impacted negatively on the market.

$$\boldsymbol{M1}: \quad M_t = \alpha + \sum_{i=1}^{3} \beta_i M_{t-i} + \sum_{i=1}^{3} {}'\Upsilon_i VOL_{t-i} + \sum_{i=1}^{3} \delta_i VML_{t-i} + \varepsilon_{1t}$$

$$\boldsymbol{M2}: \quad M_t = \alpha + \sum_{i=1}^{3} \beta_i M_{t-i} + \sum_{i=1}^{3} {}'\Upsilon_i VOL_{t-i} + \sum_{i=1}^{3} \delta_i VML_{t-i} + \sum_{i=1}^{3} \lambda_i A_{t-i} + \varepsilon_{2t}$$

Where $\mathbf{R_t}$ = $\log(SP_{t+1}) - \log(SP_t)$ | **Mt** = $R_{t+1} - R_t$ | $\mathbf{VOL_t}$ = $(R_{t+1} * R_{t+1}) – (R_t * R_t)$ | $\mathbf{VLM_t}$ = $\log(Volume_t / Volume_{t-1})$

- Upgraded the previous
  - Replaced the Monte Carlo to Monte Carlo inverse transform and a bootstrap sampling method.
  - Used a non-linear Granger causality test predictive power of the anxiety index on the market trend
- Results:
  - The theoretical and empirical *F-statistics* were still significantly apart
  - Confirmed the presence of residuals heteroscedasticity biased the prediction power

Literature Review (ctn'd)

A Statistical Approach (Ctn'd)

- Olaniyan and Al (2015) re-oriented the previous research
  - Introduced a new set of attributes:
    - Abandoned the Anxiety index and Positive and Negative sentiments attributes generated from Downside Hedge Twitter Sentiment indicator.
    - Replaced the Volatility $VOL_t = (R_{t+1} * R_{t+1}) – (R_t * R_t)$ by an EGARCH volatility (Q_t)

- Result:
  - Ljung-Box test shows positive sentiment reduces volatility but negative sentiment do not seem to have a significant impact.

  - Linear Granger causality test showed M2 outperform M1, however the experiment suffer the same autocorrelation, heteroscedasticity and non-normal distribution of the residuals as the experiment ran by Gilbert and Karahalios (2010).

  - The Monte Carlo and sampling Monte Carlo reached the same conclusion as the linear Granger causality test. However it suffered the theoretical vs empirical *F-statistics* divergence issue.

  - The non-linear Granger causality test proposed by Baek and Brock (1992) showed that sentiment had no significant impact on predicting the stock market return.

Literature Review (ctn'd)

Back to Machine Learning and NNs

- Author: Olaniyan and Al (2015)

- Methodology:
  - Attributes: Q_{t-1} , Q_{t-3} , P_{t-1} , P_{t-2} , N_{t-1} , N_{t-2}
  - Response variable: Q_t
  - Models: Feed-forward neural network (NN)|Elman recursive NN | Jordan recursive NN
  - Results:
    - Past volatility was a main contributor to predicting future volatility.
    - Positive sentiment was the main contributor to predicting future volatility .
    - Negative sentiment appeared to have less predictive power in predicting future volatility .

Limitations of current approaches

Bags-of-Words approach

- Ambiguity relative to word combination & context, e.g. ‘low quality’ vs ‘low price’
- Lexicons (SentiWordNet) sentiment likelihood limitations
- Relatively small tweet volume under analysis

The attribute Selection

- Usually a small number of technical indicators under analysis

The correlation & collinearity issue

The use of non time-series machine learning methodologies

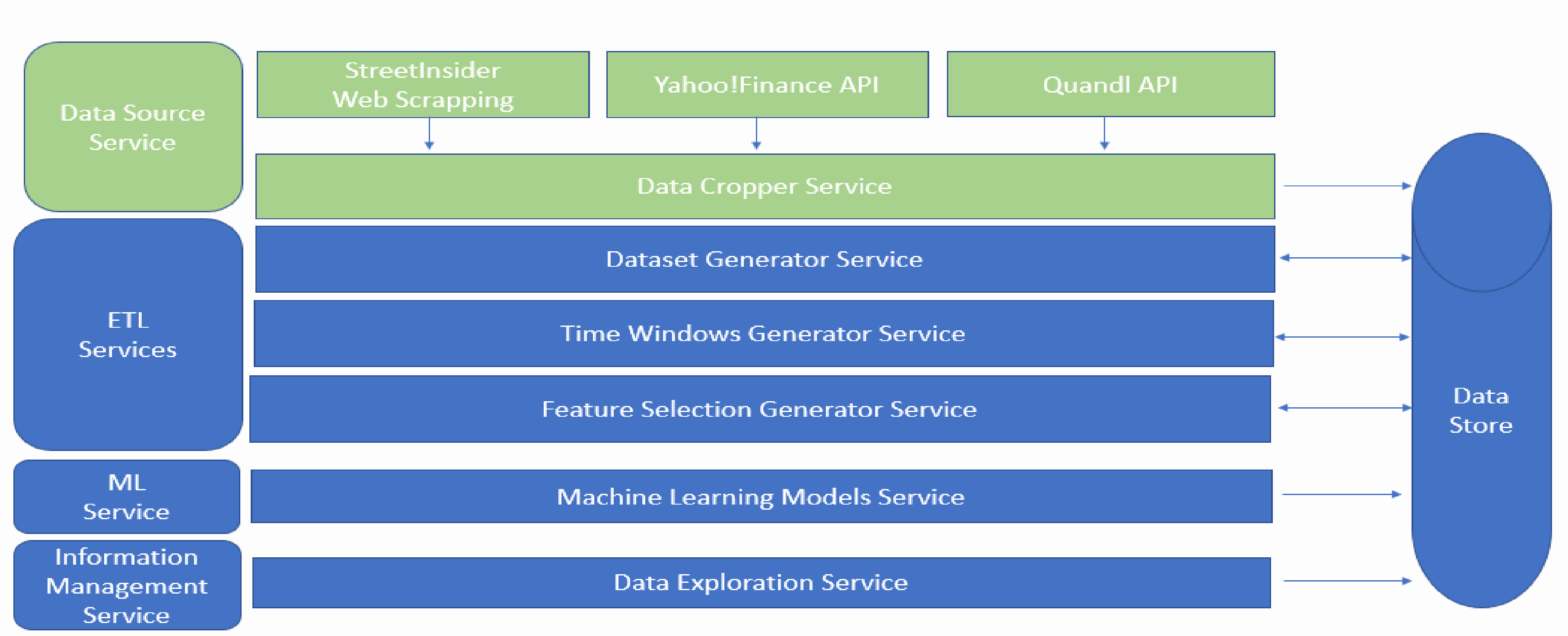
- The validation set
- The cross-validation approach

The proposed solution (in a nutshell)

High Level Description

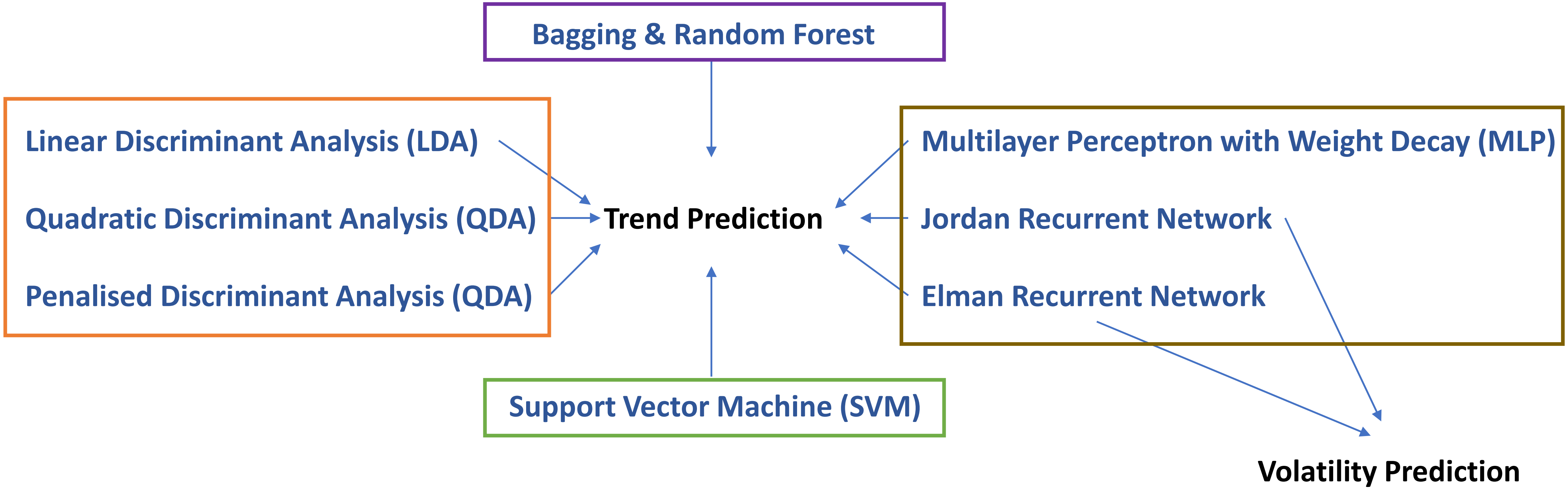
- Provide a large set of technical indicators, generating +50 explanatory variables
- Gather sentiment data an independent and a complex engine generation: *Quandl*
- Implement a robust data processing stage
- Generate a feature selection based on coupling a Wrapper and Filter method
- Apply a sliding time window
- Measure the impact of sentiment on the trend and the volatility predictability

Technical Infrastructure



The proposed solution (in a nutshell)

Machine Learning Supervised Classification/Regression Algorithms



The proposed solution (in a nutshell)

Innovation

- 50+ technical indicators under analysis
- Deployment of a "2-way" feature selection process, followed by a sliding time window for performance measurement
- Study of an entire index constituents' sentiment impact for the trend and volatility prediction

Challenges

- Ensure the prediction based on endogenous factors only were as accurate as possible.
  … 50+ market data driven indicators
  … 2-way feature selection

- No sentiment for the index.
  … Fabricated the index sentiment from the constituents' sentiment, across each stock times series)

- Missing sentiment the XLF index.
  … Moved to another index (XLE)

- The prediction power of sentiments on the trend was disappointing.
  … Used 30days, 100days and 180days for the sliding training period (keeping the validation set in the same proportion).
  … Introduced the PDA and RNNs models
  … Looked the prediction power of sentiment on the volatility

- The trend prediction kappa's were low (around 10%)
  … Low Kappa's can be recorded because of high values of concordance => used the Prevalence and Bias Adjusted Kappa, Byrt (1993).

- Three class analysis (Neutral/Up/Down) could produce misleading results.
  … Skewed the results too much => combined the neutral class into the down class.

The Methodology

Data Collection

Raw Data

- Market data downloaded from *Yahoo!Finance* API (HLCO price & Volume), over a 20 years period for most stocks.
- Sentiment data downloaded from *Quandl* , sentiment scores between -1 and +1

  Note: *Quandl* is complex engine (20 millions news article => uses deep learning + bag-of-words + n-grams)

Index sentiment Generation

- *Quandl* does not provide index sentiment, a proxy was built from the constituents' sentiment

$SIS_t = \sum_{i=1}^{n}(SS_i * W_i)_t$, where n is the number of stocks

Data Generation

Trend Prediction (supervised classification problem)

- Response Variable:
  - $R_t = \log(Close_{t+1} / Close_t)$
  - Dummification: $\begin{cases} Return_t > 0 \text{ then } Direction\ is\ set\ to\ Up \\ Return_t \leq 0 \text{ then } Direction\ is\ set\ to\ Down \end{cases}$
- Explanatory Variables:
  - $Close_t$ and its lags / $Volume_t$ and its lags
  - A Suite of technical indicators, e.g. ROC, SMA, Momemtum, RSI, etc.

Volatility Prediction (supervised regression problem)

- Response Variable:
  - The volatility proxy r^2
- Explanatory Variables:
  - EGARCH $volatility_t$ and its lags
  - $Volume_t$ and its lags

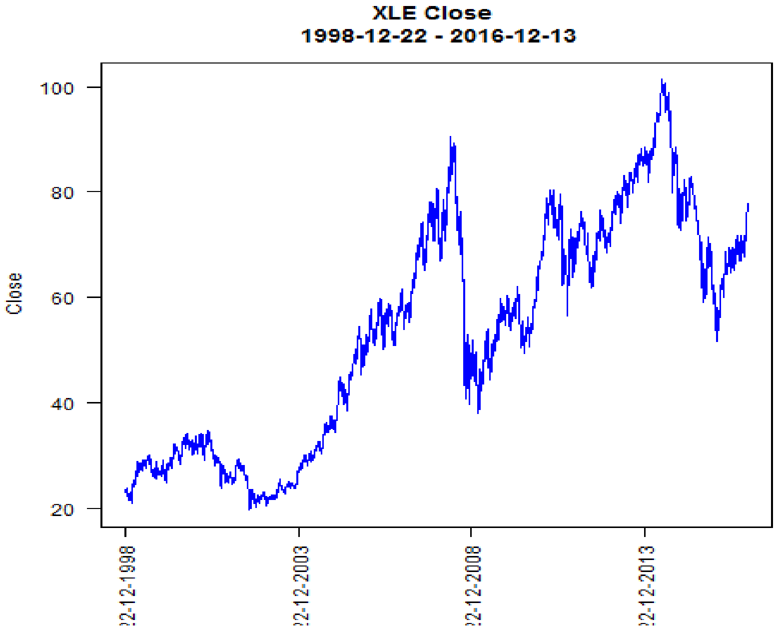
The Methodology

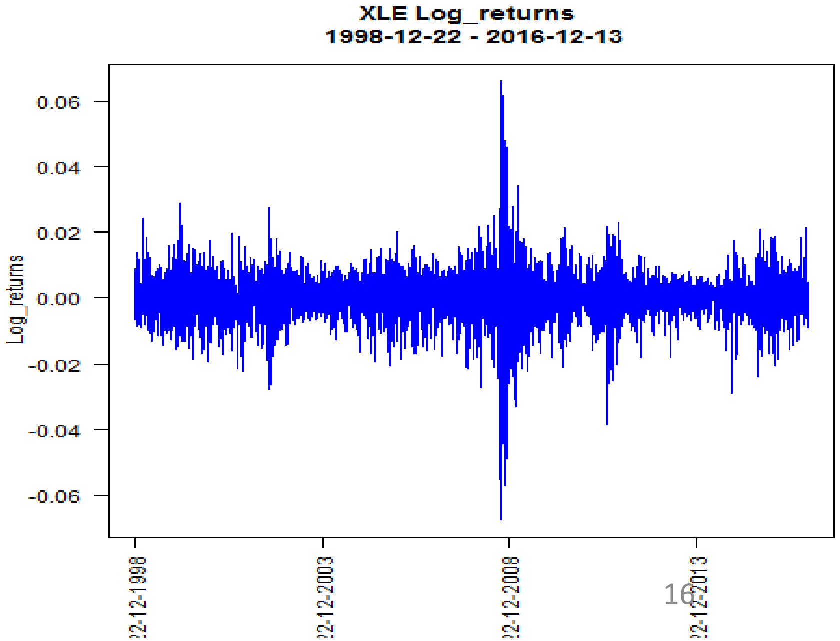
The Choice of the XLE index

- Originally started with the XLF index but sentiment data was missing
    - SPGI (S&P Global Inc) and WLTW (Willis Towers Watson PLC) had no sentiment
    - BRK-B (Berkshire Hathaway B) - the 1st largest weight (index weight = 10%) -> missing 85% of the sentiment data.
    - BAC (Bank of America Corp) - the 4th largest weight, (index weight = 8%) -> missing 37% of the sentiment data.
- XLE is a better fit:
    - All constituents have sentiment information
    - The first 2 constituents, which represent 17%, 15% of the index were only missing 6%, 2% of the sentiment data.

The Explanatory Data Analysis

- All explanatory values are continuous
- We are dealing with time series





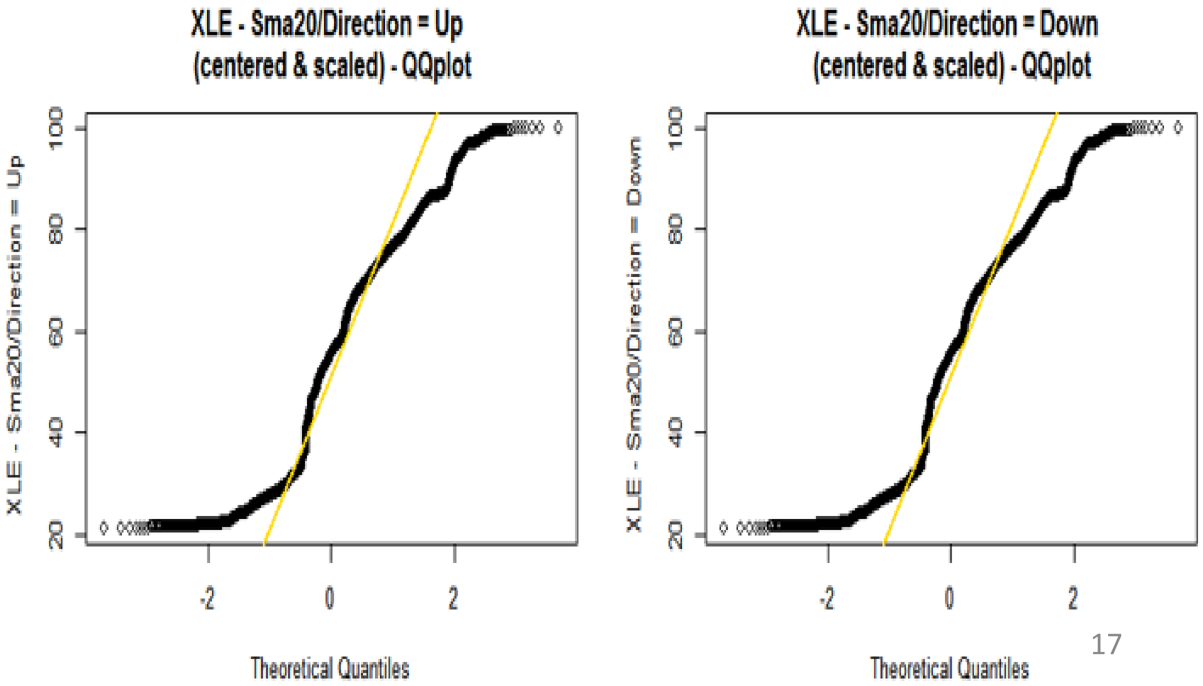
The Methodology

The Explanatory Data Analysis (ctn'd)

- Response variable
  - Most assets show a negative skew –> sign of asymmetry from the ND
  - Most assets show a Kurtosis > 3 –> leptokurtic distribution with thicker tails

Name	Minimum	Maximum	Mean	Median	Variance	Skewness	Kurtosis
XLE	−0.067748	0.066231	0.000114	0.000274	0.000058	−0.399458	8.570529
APA	−0.311733	0.083923	0.000058	0	0.000129	−2.646812	70.680149
APC	−0.308981	0.092018	0.000068	0	0.000133	−5.019408	131.064398
BHI	−0.197411	0.108233	0.000072	0	0.00013	−0.69397	17.957388

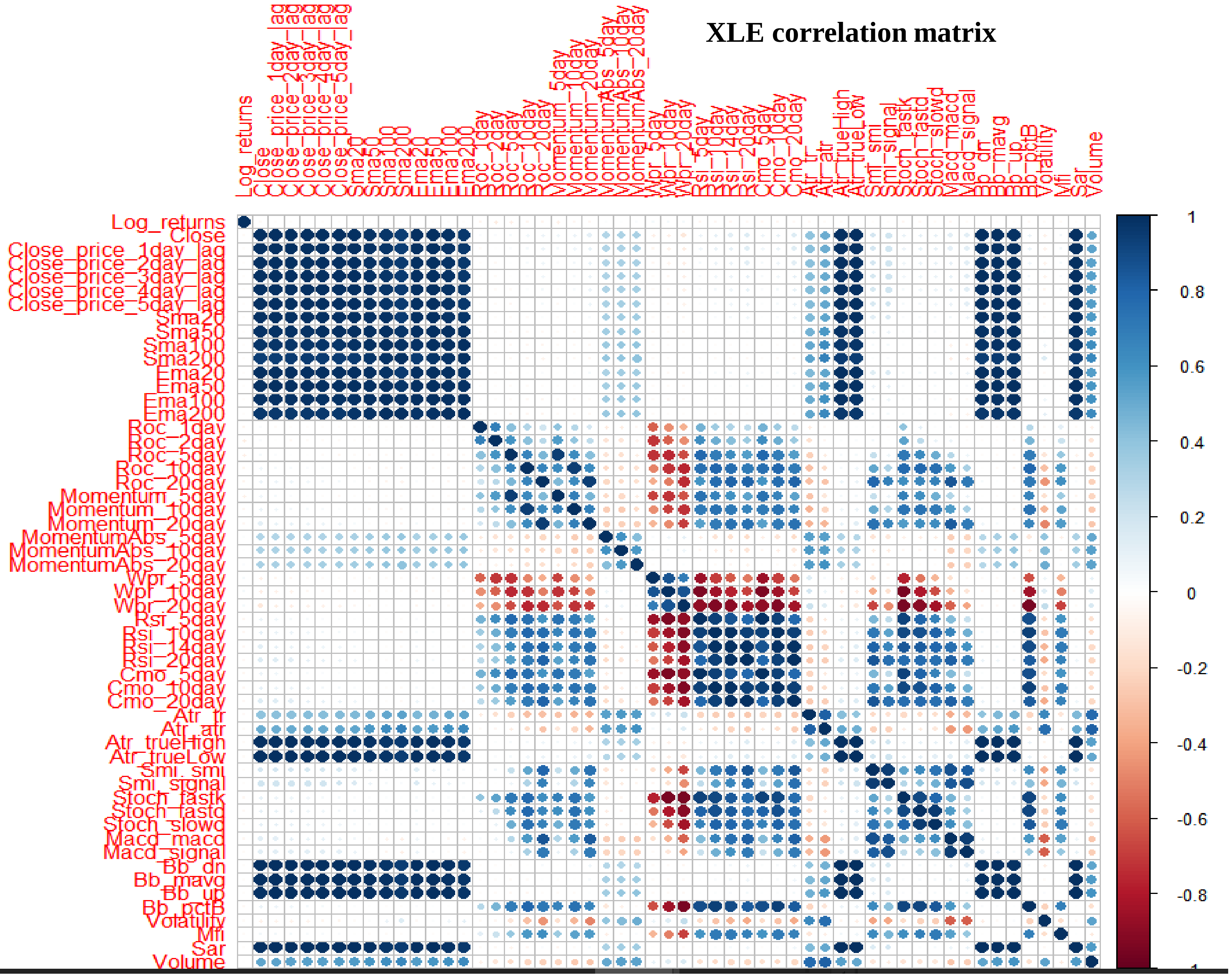
- Explanatory variables
  - Kolmogorov–Smirnov test *Null* hypothesis (H0) indicating the data distribution seems to follow a ND is rejected most of time.



The Methodology

The Explanatory Data Analysis (ctn’d)

- Correlation & Multicollinearity
    - Very small degree of correlation between the explanatory and response variables
    - High degree of correlation between some of the explanatory variables (e. g. SMA20 and SAR)



The Methodology

Pre-processing

- Missing Data
  - Market Data
    - Technical indicators lags generate missing data
    - Remediation: removal
  - Sentiment
    - Data can be missing for a few consecutive days (ex: APA between 2 and 10 consecutive days)
    - Remediation: median imputation

- Generic Step removes:
  - Near zero variance columns
  - Linearly dependent columns
  - Attributes showing a correlation within themselves greater than 95%.

- Model Specific Step:
  - Apply Box-Cox transform for models that require the attributes ND (e.g. LDA)
  - Same comment for Scaling/Centring (e.g. SVM)

The Methodology

Pre-processing (Ctn'd)

- Class-rebalancing:

Frequency	returns = 0	returns > 0	returns < 0
Mean	5.34%	47.88%	46.78%
Median	5.96%	47.63%	46.41%
Std Dev	3.27%	2.31%	1.40%
Min	0.29%	42.79%	44.17%
Max	12.96%	53.37%	50.65%

Table 3 – Descriptive statistics for a three classes response variable

Frequency	returns > 0	returns <= 0
Mean	47.88%	52.12%
Median	47.63%	52.37%
Std Dev	2.31%	4.67%
Min	42.79%	44.46%
Max	53.37%	63.61%

Table 4 – Descriptive statistics for an aggregated two classes response variable

- No SMOTE -> it reshuffles the data
- No epsilon -> migrate too many positive/negative returns towards the 0 returns

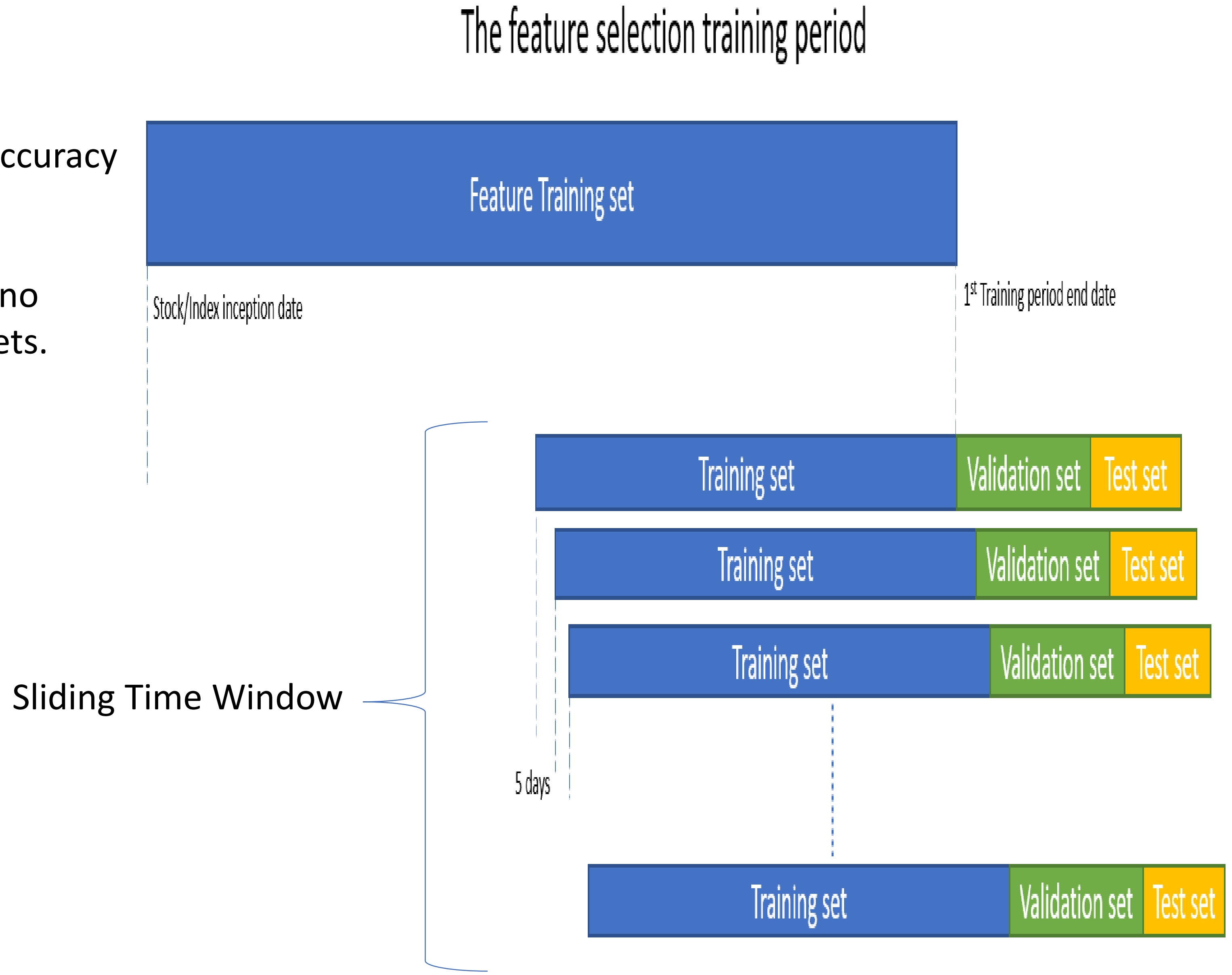
=> Instead migration of the 0 returns towards the negative bucket.

The Methodology

Feature Selection

- The aim is to obtain the ‘best’ base line accuracy rate for each model under analysis.

- The training period is defined so there is no overlap with the validation or test data sets.



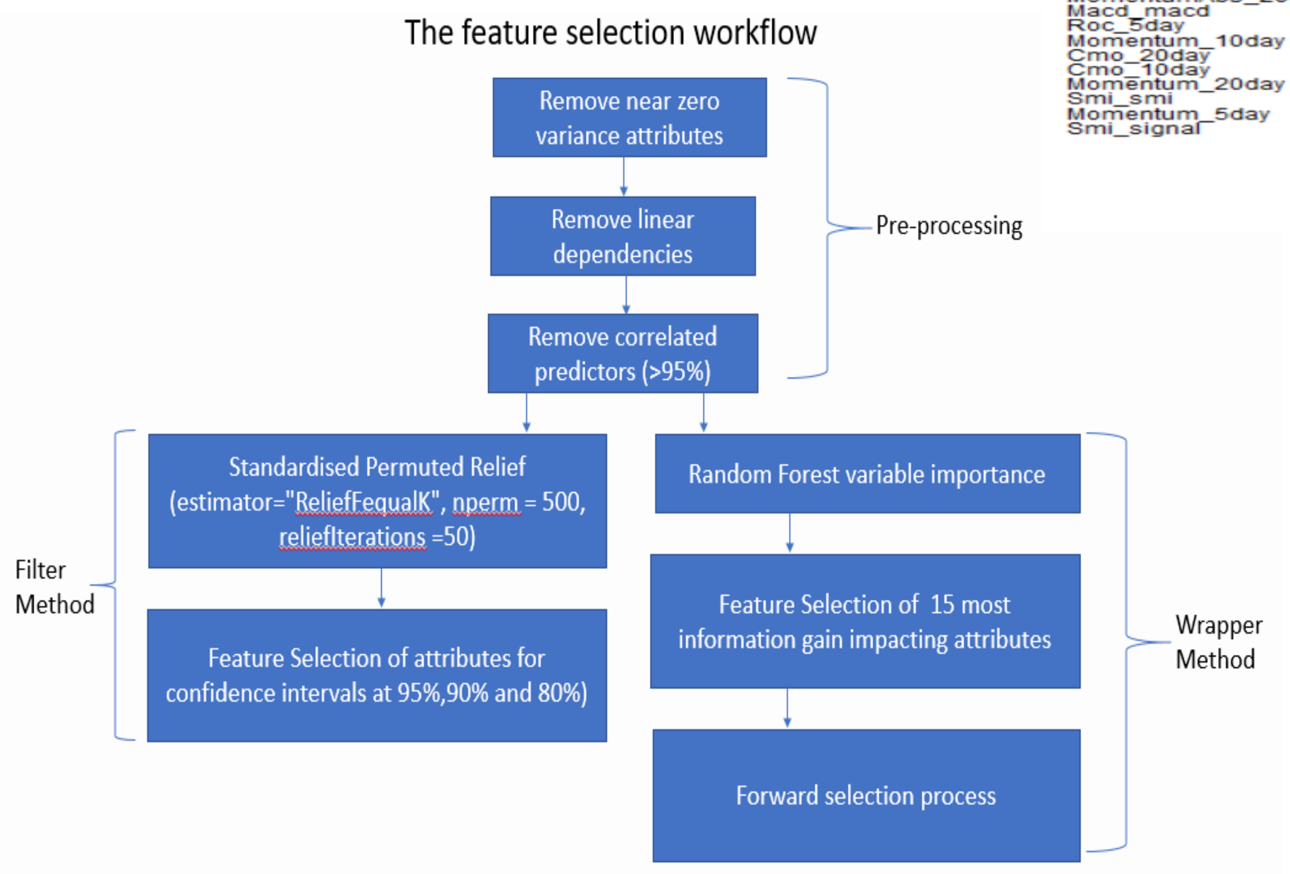
The Methodology

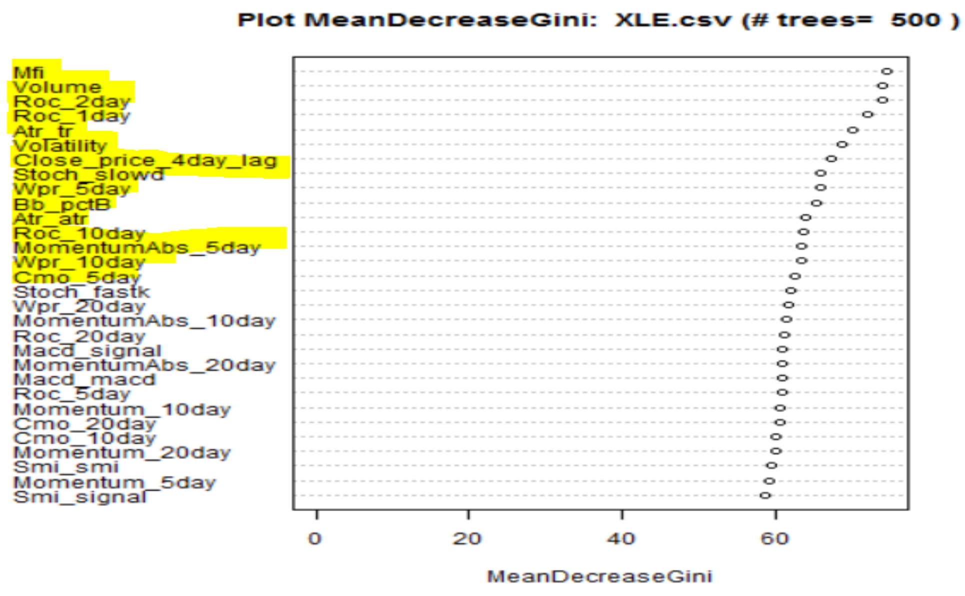
Feature Selection for the trend (cnt’d)

- Implementation of a “2-way” feature selection, using a Wrapper and a Filter method.

	perm.standardized
Momentum_10day	2.074645265
MomentumAbs_10day	1.793285857
Roc_10day	1.448368777
Smi_smi	1.398109204
MomentumAbs_5day	1.084293572
Wpr_5day	1.056260694
Smi_signal	0.929895318
Momentum_5day	0.860893427
Roc_20day	0.842843114
Bb_pctB	0.667610504
Stoch_slowd	0.362669476
Stoch_fastk	0.278135083
Roc_5day	0.105564989
Cmo_5day	-0.062617825
Roc_1day	-0.124363492
Wpr_20day	-0.198234982
Macd_macd	-0.458359156
Mfi	-0.48361069
Wpr_10day	-0.601287442
Momentum_20day	-0.906115256
Atr_atr	-0.942968961
Atr_tr	-1.005896359
MomentumAbs_20day	-1.03254572
Close_price_4day_lag	-1.281339064
Macd_signal	-1.432551332
Volatility	-1.692683785
Cmo_10day	-2.226379389
Roc_2day	-2.534294844
Cmo_20day	-2.545368963
Volume	-2.823705385

XLE (|level|>1.96 – 95% c.i.)





formula1 = Mfi

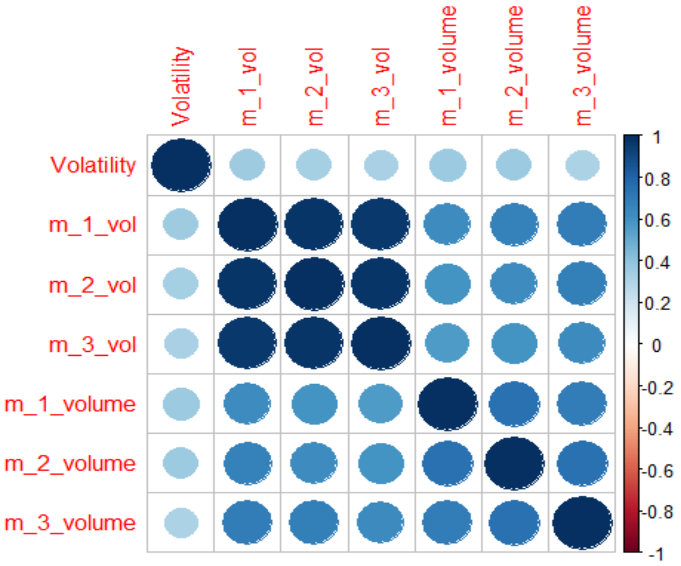
formula2 = Volume + Roc_2day

…

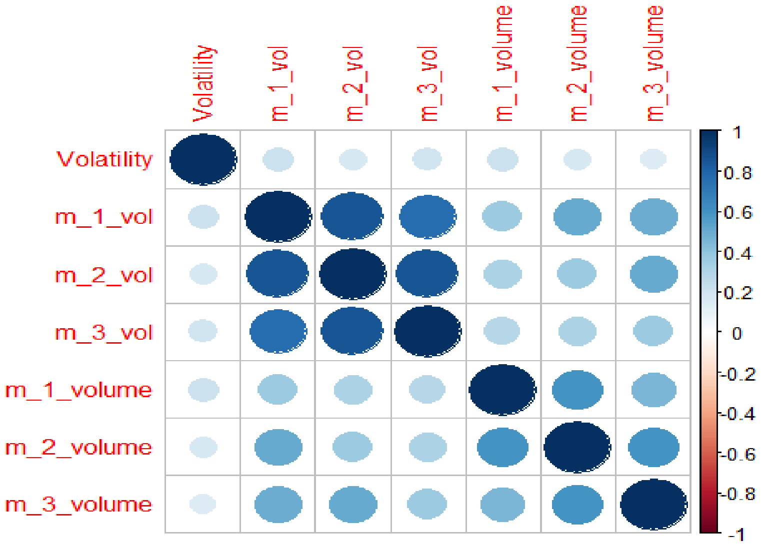
formula15 = Mfi+ Volume + …. + Cmo_5day

The Methodology

Feature Selection for the volatility (cnt'd)



XLE index correlation matrices (Volatility is the response variable)



EOG stock correlation matrices (Volatility is the response variable)

1. XLE index: high degree of correlation between the volatility $GARCH_{t-1}$ and $GARCH_{t-2}$ (98%) & the volatility $GARCH_{t-1}$ and $GARCH_{t-3}$ (96%). The volatility $GARCH_{t-1}$ is selected, the two others are dropped.

2. EOG Stock: high degree of correlation between the volatility and its respective lags at respectively 85% and 76%. But, the volatility $GARCH_{t-2}$ and $GARCH_{t-3}$ are retained, as they are below the 95% correlation cut off.

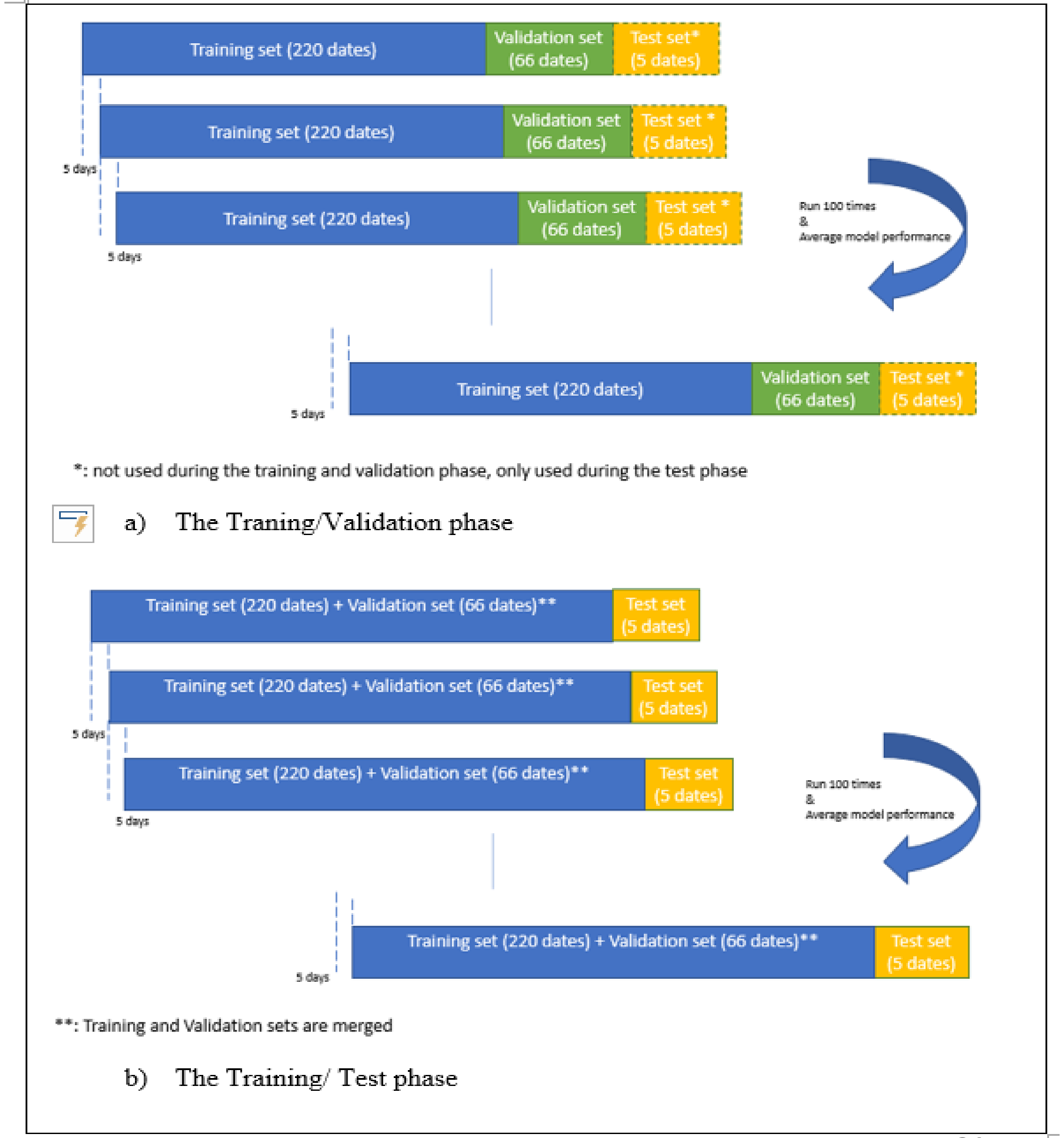
The volume and its lags, for both groups, do not show high level of multi-collinearity. Therefore, the lagged volume variables are all selected.

The Methodology

The Sliding Time Window

- A time window is composed of a contiguous 100 time-slices of equal lengths.

- Each time slide is divided in three parts:
  - A training set containing 220 records,
  - A validation set holding the next 66 dates (30% of the training data), and
  - A test set representing 5 business days of data.



a) The Traning/Validation phase

b) The Training/ Test phase

The Methodology

The Results Generation Framework (Trend)

- The models and their parameters

Model Name	*Hyper-parameters*
Logistic Discriminant Analysis (LDA)	None
Quadratic Discriminant Analysis (QDA)	None
Penalised Discriminant Analysis (PDA)	None
Support Vector Machine (SVM)	Kernel = Linear Cost = (0.001, 0.01, 0.1, 1 ,100)
Random Forest	Tree number = 500 Random Forest splits = $\sqrt{p}$ or $\frac{mp}{2}$. Bagging splits = p *'p'* represents the number of attributes in the model.
Multilayer Perceptron with Weighted Decay (MLP)	Hidden Layer size = [1,2,…,15] The weight decay = (0,0.001,0.1,1) Number of iterations = 1000
Elman and Jordan Recursive Neural Networks (RNN)	Hidden Layer size = (5,7,10,15,20) The weight decay = (0,0.001,0.01,0.1,1) Number of iterations = (100,500,1000,1500,2000)

- The base scenarios generation for each asset:

A_{ij} = Max (Max Wrapper Test Accuracy $Rate_{ij}$, Filter Test Accuracy $Rate_{ij}$)

Where:
Max Wrapper Accuracy $Rate_{ij}$ =
Max ($Wrapper_1$ Test Accuracy $Rate_i$,…, $Wrapper_k$ Test Accuracy $Rate_i$ $)_j$
i represents the i^{th} asset
k represents the k^{th} Feature Selection list
j represents the j^{th} Machine Learning model

- The sentiment scenarios:
  - Measure the impact of sentiment S_t, S_{t-1} S_{t-2} S_{t-3} independently
  - Measure the impact of the sentiment momentum SM_t, SM_{t-1} SM_{t-2} SM_{t-3} independently, where $SM_t = S_t - S_{t-1}$

The Methodology

The Results Generation Framework (Volatility)

- The models and their parameters

Elman and Jordan Recursive Neural Networks (RNN)	Hidden Layer size = (5,7,10,15,20) The weight decay = (0,0.001,0.01,0.1,1) Number of iterations = (100,500,1000,1500,2000)

- The base scenarios generation for each asset:

A_i = Max (Test Accuracy Rate$_{i\ Jordan.}$ Test Accuracy Rate$_{i\ Elman.}$)

Where:
i represents the i^{th} asset

- The sentiment scenarios:
  - Measure the impact of sentiment S_{t-1} independently
  - Measure the impact of the sentiment momentum SM_{t-1} and SM_{t-2} independently, where $SM_t = S_t - S_{t-1}$

The Methodology

Performance measures

- Trend Prediction
    - Accuracy Rate
    - Kappa

- Volatility Prediction
    - MSE
    - RMSE

The Results

The Impact of Sentiment on the Trend

- The below table presents the summary Test Accuracy rates for each scenario, when
  i) the sentiment data is added at the index level,
  ii) the sentiment data is tested at the constituents level

	Base Scenario	St	St-1	St-2	St-3	SMt	SMt+SMt-1	SMt+SMt-1+SMt-2	SMt+SMt-1+SMt-2+SMt-3
Index	54%	53%	50%	49%	52%	49%	45%	48%	48%
Sum of Weigthed Constituents	54%	53%	53%	53%	53%	53%	52%	52%	52%

The Impact of Sentiment on the Volatility

- The below table presents the summary RMSE for each scenario, when
  i) the sentiment data is added at the index level

	Base Scenario	St-1	SMt+SMt-1
Index	0.002010	0.000650	0.000691

ii) the sentiment data is tested at the constituents level.
  - the figures in the table below correspond to the Total Weighted RMSE

	St-1	SMt+SMt-1
Constituents	0.002900	0.000319

The Results

The Impact of Sentiment on the Volatility (in details)

code	model_name	scenario	type	mse	rmse	type	mse	rmse	delta S1 - Base	weights	weighted rmse
EOG	jordan	sentiment_m1	validation	1.22428E-06	0.001106473	test	3.69082E-06	0.001921151	-0.000545374	4.64%	8.91414E-05
HAL	jordan	sentiment_m1	validation	5.98251E-07	0.000773467	test	3.04904E-07	0.000552181	-0.000142788	3.66%	2.02098E-05
NBL	jordan	sentiment_m1	validation	2.86618E-06	0.00169298	test	1.88432E-06	0.001372707	-0.000446403	1.58%	2.16888E-05
OXY	jordan	sentiment_m1	validation	4.59378E-07	0.000677774	test	1.72062E-06	0.001311725	-0.000201332	3.14%	4.11882E-05
APA	elman	sentiment_m1	validation	5.40497E-06	0.002324859	test	2.0349E-05	0.004510981	0.000627052	1.91%	8.61597E-05
APC	elman	sentiment_m1	validation	6.73796E-06	0.002595758	test	2.57376E-06	0.001604293	-0.001560004	2.98%	4.78079E-05
BHI	elman	sentiment_m1	validation	1.60363E-06	0.001266343	test	1.66168E-06	0.001289063	-0.001113514	2.43%	3.13242E-05
CHK	elman	sentiment_m1	validation	0.000457501	0.021389264	test	0.000733947	0.027091447	0.002328003	0.47%	0.00012733
COG	elman	sentiment_m1	validation	8.49214E-06	0.002914128	test	1.72927E-05	0.00415845	0.00301316	1.52%	6.32084E-05
COP	elman	sentiment_m1	validation	3.38591E-06	0.001840085	test	5.30372E-06	0.002302981	-7.46657E-05	3.12%	7.1853E-05
CVX	elman	sentiment_m1	validation	4.28324E-07	0.000654464	test	6.02729E-07	0.000776356	-0.001734916	14.81%	0.000114978
CXO	elman	sentiment_m1	validation	6.12021E-06	0.002473906	test	1.44432E-06	0.001201798	-0.002462158	1.30%	1.56234E-05
DVN	elman	sentiment_m1	validation	1.2753E-05	0.003571139	test	2.92193E-05	0.005405485	-3.64848E-05	1.88%	0.000101623
EQT	elman	sentiment_m1	validation	7.84349E-06	0.002800624	test	5.06125E-06	0.002249722	-0.00148149	0.79%	1.77728E-05
FTI	elman	sentiment_m1	validation	9.86438E-07	0.000993196	test	8.30923E-06	0.002882574	-0.000295896	0.94%	2.70962E-05
HES	elman	sentiment_m1	validation	9.19988E-06	0.00303313	test	1.88909E-06	0.001374441	-0.002603408	1.40%	1.92422E-05
HP	elman	sentiment_m1	validation	2.34418E-06	0.001531072	test	1.41077E-06	0.001187758	-0.002424238	0.58%	6.889E-06
KMI	elman	sentiment_m1	validation	1.44327E-05	0.003799035	test	1.20192E-05	0.003466878	-0.000804663	2.65%	9.18723E-05
MPC	elman	sentiment_m1	validation	6.4679E-06	0.002543206	test	4.9705E-05	0.007050176	0.004475356	1.70%	0.00011985
MRO	elman	sentiment_m1	validation	2.00918E-05	0.00448239	test	2.4106E-05	0.004909789	-0.000398091	1.20%	5.89175E-05
MUR	elman	sentiment_m1	validation	7.29311E-06	0.002700575	test	0.000169687	0.013026387	0.002942257	0.48%	6.25267E-05
NFX	elman	sentiment_m1	validation	9.02727E-06	0.003004541	test	3.10356E-05	0.005570959	0.004446519	0.60%	3.34258E-05
NOV	elman	sentiment_m1	validation	1.77189E-06	0.001331125	test	2.00076E-05	0.004472984	0.000917506	1.25%	5.59123E-05
OKE	elman	sentiment_m1	validation	1.8283E-05	0.004275858	test	3.20359E-05	0.005660024	0.004332364	0.80%	4.52802E-05
PSX	elman	sentiment_m1	validation	5.39742E-07	0.000734672	test	1.5293E-07	0.000391063	-0.001629952	2.55%	9.9721E-06
PXD	elman	sentiment_m1	validation	2.62789E-06	0.001621078	test	1.20019E-06	0.001095532	-0.001580343	4.78%	5.23664E-05
RIG	elman	sentiment_m1	validation	3.20358E-06	0.001789855	test	0.000148241	0.012175436	0.002415273	0.37%	4.50491E-05
RRC	elman	sentiment_m1	validation	1.85391E-05	0.00430571	test	1.73119E-05	0.004160751	0.000112819	0.68%	2.82931E-05
SE	elman	sentiment_m1	validation	1.79938E-06	0.00134141	test	2.43E-07	0.000492951	-0.001745335	2.53%	1.24717E-05
SLB	elman	sentiment_m1	validation	4.22883E-07	0.000650295	test	6.57553E-07	0.000810897	-0.001865363	8.19%	6.64124E-05
SWN	elman	sentiment_m1	validation	5.28955E-05	0.007272931	test	0.000185669	0.013626054	0.001534635	0.46%	6.26798E-05
TSO	elman	sentiment_m1	validation	5.51196E-06	0.002347757	test	1.16883E-06	0.001081126	-0.001389606	2.22%	2.4001E-05
VLO	elman	sentiment_m1	validation	2.50129E-06	0.001581547	test	3.19487E-07	0.000565232	-0.001734745	2.84%	1.60526E-05
WMB	elman	sentiment_m1	validation	0.000632759	0.025154703	test	0.003792647	0.061584471	0.05794406	1.87%	0.00115163
XEC	elman	sentiment_m1	validation	3.88223E-06	0.001970337	test	1.11761E-06	0.001057173	-0.001701143	0.86%	9.09169E-06
XLE	elman	sentiment_m1	validation	3.49233E-07	0.00059096	test	4.1791E-07	0.000646459	-0.001367147		0
XOM	elman	sentiment_m1	validation	2.86713E-07	0.000535456	test	9.39536E-08	0.000306519	-0.00196712	16.80%	5.14951E-05
										total	0.002900439

code	model_name	scenario	type	mse	rmse	type	mse	rmse	delta SMM1M2 - Base	weights	weighted rmse
EOG	jordan	sentiment_momentum_mm1m2	validation	1.26247E-06	0.001123596	test	3.88143E-06	0.001970135	-0.00049639	4.64%	9.14142E-05
HAL	jordan	sentiment_momentum_mm1m2	validation	6.16404E-07	0.000785114	test	2.48947E-07	0.000498946	-0.000196023	3.66%	1.82614E-05
NBL	jordan	sentiment_momentum_mm1m2	validation	2.99022E-06	0.001729226	test	1.4098E-06	0.001187348	-0.000631762	1.58%	1.87601E-05
OXY	jordan	sentiment_momentum_mm1m2	validation	4.42033E-07	0.000664855	test	1.73887E-06	0.001318663	-0.000194393	3.14%	4.1406E-05
APA	elman	sentiment_momentum_mm1m2	validation	5.38433E-06	0.002320416	test	1.13616E-05	0.00337069	-0.000513239	1.91%	6.43802E-05
APC	elman	sentiment_momentum_mm1m2	validation	8.08767E-06	0.002843883	test	6.09037E-06	0.002467868	-0.000696429	2.98%	7.35425E-05
BHI	elman	sentiment_momentum_mm1m2	validation	6.18371E-06	0.002486706	test	9.95769E-06	0.00315558	0.000753004	2.43%	7.66806E-05
CHK	elman	sentiment_momentum_mm1m2	validation	0.000411966	0.020296941	test	0.000547635	0.02340161	-0.001361834	0.47%	0.000109988
COG	elman	sentiment_momentum_mm1m2	validation	1.77473E-05	0.004212751	test	2.63232E-05	0.005130615	0.003985325	1.52%	7.79854E-05
COP	elman	sentiment_momentum_mm1m2	validation	3.79659E-06	0.001948484	test	3.61728E-06	0.001901915	-0.000475732	3.12%	5.93397E-05
CVX	elman	sentiment_momentum_mm1m2	validation	2.87533E-06	0.00169568	test	4.8817E-06	0.002209457	-0.000301816	14.81%	0.000327221
CXO	elman	sentiment_momentum_mm1m2	validation	6.31272E-06	0.002512512	test	1.22911E-05	0.003505862	-0.000158095	1.30%	4.55762E-05
DVN	elman	sentiment_momentum_mm1m2	validation	1.34492E-05	0.003667317	test	4.3922E-05	0.00662737	0.0011854	1.88%	0.000124595
EQT	elman	sentiment_momentum_mm1m2	validation	1.24705E-05	0.003531365	test	1.31329E-05	0.003623933	-0.000107279	0.79%	2.86291E-05
FTI	elman	sentiment_momentum_mm1m2	validation	3.68056E-06	0.001918478	test	7.06667E-06	0.002658321	-0.000520148	0.94%	2.49882E-05
HES	elman	sentiment_momentum_mm1m2	validation	9.80629E-06	0.0031315	test	2.65977E-05	0.005157299	0.00117945	1.40%	7.22022E-05
HP	elman	sentiment_momentum_mm1m2	validation	2.98943E-06	0.001728996	test	5.43633E-06	0.002331594	-0.001280401	0.58%	1.35232E-05
KMI	elman	sentiment_momentum_mm1m2	validation	1.49277E-05	0.003863636	test	1.77158E-05	0.004209018	-6.25239E-05	2.65%	0.000111539
MPC	elman	sentiment_momentum_mm1m2	validation	5.11662E-06	0.002261995	test	1.13276E-05	0.003365655	0.000790835	1.70%	5.72161E-05
MRO	elman	sentiment_momentum_mm1m2	validation	1.69012E-05	0.004111112	test	0.000403828	0.020095481	0.014787602	1.20%	0.000241146
MUR	elman	sentiment_momentum_mm1m2	validation	9.04276E-06	0.003007118	test	0.000158962	0.012608016	0.002523885	0.48%	6.05185E-05
NFX	elman	sentiment_momentum_mm1m2	validation	3.19018E-05	0.005648167	test	4.9306E-05	0.007021826	0.005897386	0.60%	4.2131E-05
NOV	elman	sentiment_momentum_mm1m2	validation	3.43824E-06	0.001854249	test	2.31884E-05	0.004815432	0.001259954	1.25%	6.01929E-05
OKE	elman	sentiment_momentum_mm1m2	validation	1.90792E-05	0.004367976	test	1.09973E-05	0.003316214	0.001988554	0.80%	2.65297E-05
PSX	elman	sentiment_momentum_mm1m2	validation	2.9163E-06	0.001707718	test	5.36014E-06	0.002315198	0.000294183	2.55%	5.90375E-05
PXD	elman	sentiment_momentum_mm1m2	validation	3.44838E-06	0.001856981	test	1.45162E-05	0.003810008	0.001134133	4.78%	0.000182118
RIG	elman	sentiment_momentum_mm1m2	validation	4.62669E-06	0.002150975	test	0.000103641	0.010180426	0.000420264	0.37%	3.76676E-05
RRC	elman	sentiment_momentum_mm1m2	validation	1.87136E-05	0.004325922	test	8.03385E-06	0.002834404	-0.001213528	0.68%	1.9274E-05
SE	elman	sentiment_momentum_mm1m2	validation	2.86831E-06	0.00169361	test	2.50624E-06	0.00158311	-0.000655176	2.53%	4.00527E-05
SLB	elman	sentiment_momentum_mm1m2	validation	6.43362E-06	0.002536457	test	8.2859E-06	0.002878524	0.000202265	8.19%	0.000235751
SWN	elman	sentiment_momentum_mm1m2	validation	5.86402E-05	0.007657689	test	0.000171642	0.013101225	0.001009806	0.46%	6.02656E-05
TSO	elman	sentiment_momentum_mm1m2	validation	7.49381E-06	0.002737483	test	7.90498E-06	0.002811579	0.000340847	2.22%	6.24171E-05
VLO	elman	sentiment_momentum_mm1m2	validation	2.47586E-06	0.001573486	test	3.37726E-06	0.001837732	-0.000462245	2.84%	5.21916E-05
WMB	elman	sentiment_momentum_mm1m2	validation	0.000663628	0.025760971	test	9.93145E-05	0.009965668	0.006325257	1.87%	0.000186358
XEC	elman	sentiment_momentum_mm1m2	validation	5.10921E-06	0.002260357	test	8.52769E-06	0.00292022	0.000161905	0.86%	2.51139E-05
XLE	elman	sentiment_momentum_mm1m2	validation	5.47508E-07	0.000739938	test	4.77191E-07	0.00069079	-0.001322816		0
XOM	elman	sentiment_momentum_mm1m2	validation	2.62116E-06	0.001619	test	4.67214E-06	0.002161513	-0.000112126	16.80%	0.000363134
										total	0.003191146

Conclusion

☹

- Sentiment and sentiment momentum do not seem to have a positive impact on the index **trend prediction**.
  - This is in disagreement with the machine learning literature, e.g. Parikh and Shah (2015) and Halgamuge (2007) but in agreement with the statistically more robust approach offered by Gilbert and Karahalios (2010) and Olaniyan R. et al (2015).
  - It should be noted that our machine learning approach is more robust and complete (use of sliding window, use of a strong feature selection methodology, etc.) than the one proposed in the literature.

☺

- Sentiment and sentiment momentum do seem to have a positive impact on the index **volatility prediction**.
  - This is in line with the findings from Olaniyan R. et al (2015).

Recommendations

- Generate the predicted index volatility and compare the different scenarios to establish which sentiment and/or sentiment momentum generates the best prediction.

$$\sigma_w^2 = w^T S w = [w_1, \ldots, w_2] \begin{bmatrix} \sigma_{11} & \ldots \ldots & \sigma_{1N} \\ \sigma_{21} & \ldots \ldots & \sigma_{2N} \\ \ldots & \ldots \ldots \ldots & \ldots \\ \sigma_{1N} & \ldots \ldots & \sigma_{NN} \end{bmatrix} \begin{bmatrix} w_1 \\ \ldots \\ \ldots \\ w_N \end{bmatrix}$$

where $w_1, \ldots, w_2$ are the weights and $\sigma_{11} \ldots \sigma_{1N}$ are the volatilities.

- Investigate the impact of sentiment with other GARCH models such as heteroskedasticity(GARCH), the Threshold ARCH (TARCH), the asymmetric power ARCH (APARCH) or the nonlinear GARCH (Brownlees, 2012).

- Study the impact sentiment in a the stochastic Backpropagation through time settings (Wang et Al , 2016).

- Perform a more in-depth investigation on the S&P500 to confirm/inform the current results on the trend prediction.

- Implement a time varying index sentiment proxy. But this has a cost…

- Generate missing *Quandl* sentiment

References (ctn'd)

Abdulkarim S.A. 2016. "Time Series Prediction with Simple Recurrent Neural Networks". *Bayero Journal of Pure and Applied Sciences.* 9(1): pp.19–24.

Accern. "Alpha One Guide Book – The most comprehensive trading analytics data set, Actionable Trading Analytics". http://joshua.mcelfre.sh/AlphaOne_UserGuide.pdf (accessed February 1, 2009).

Accern. "Alphaone News Sentiment". https://www.quantopian.com/data/accern/alphaone. (Accessed 12 Jun 2017).

Altman, D.G. 1991. *Practical Statistics for Medical Research.* Chapman & Hall. London.

Ang A., Timmerman A. (2011), "Regime Changes and Financial Markets". *Netspar Discussion Papers*. DP–06/20011–068. pp1–19.

Brownlees C., Engle R., and Kelly B. (2012), "A practical guide to volatility forecasting through calm and storm", *The Journal of Risk*. Volume 14/Number 2. pp3-22

Bachelier L. 1900, "The Theory of Speculation". http://www.radio.goldseek.com/bachelier–thesis–theory–of–speculation–en.pdf. (Accessed 12 Jul 2017).

Baek E., and Brock W. 1992. "A general test for nonlinear Granger causality: bivariate model". Working paper. Iowa State University.

Chawla NV., Bowyer KW, Hall LO., Kegelmeyer WP.2002. "SMOTE: Proxy Minority Over-Sampling Technique". *Journal of Artificial Intelligence Research*. pp321-357.

Cohen J. (1960). "A Coefficient of Agreement for Nominal Scales". *Educational and Psychological Measurement*. pp.20-37.

Fama E. 1964. "The behavior of stock market prices". *Journal of Business*. Volume 38, Issue 1. p34–105.

Frank J. Massey Jr. 1951. "The Kolmogorov–Smirnov Test for Goodness of Fit." *Journal of the American Statistical Association*. Vol. 46, No. 253. pp. 68– 78

Gilbert E., and Karahalios K 2010. "Widespread worry and the stock market". *In Proceedings of the 4th International Conference on Weblogs and Social Media*.pp58–65.

Giles D.E. 2007. "Some Properties of Absolute Returns as a Proxy for Volatility", *Econometrics Working Paper EWP0706*, University of Victoria, ISSN 1485-6441, p3.

Granger C.W.J. 1969. "Investigating Causal Relations by Econometric Models and Cross–spectral Methods". *Econometrica*. Vol. 37, No. 3. pp. 424–438.

References (ctn'd)

Grishman 2014. "Natural Language Processing". http://www.cs.nyu.edu/courses/spring14/CSCI–GA.2590–001/Lecture2.html *(Accessed: April 2014).*

Halgamuge S.K. 2007. "Combining News and Technical Indicators in Daily Stock Price Trends Prediction". *Conference: Advances in Neural Networks – ISNN 2007*. 4th International Symposium on Neural Networks. ISNN 2007 Proceedings. Part III.

Hastie T., Bujas A., and Tibshirani R. 1995. "Penalized Discriminant Analysis", *The Annals of Statisitcs*. Vol. 23. No 1. pp.73–192.

Haykin S. 1999, *Neural Networks – A Comprehensive Foundation*. Second Edition. Printice Hall International Inc. New Jersey. pp156–247.

Henkel S.J., Martin J.S., and Nardari F. 2011. "Time–varying Short–Horizon Predictability". *Journal of Financial Economics*, 99, pp.560–580.

Imandoust S.B., and Bolandraftar M. 2014. "Forecasting the direction of stock market index movement using three data mining techniques: the case of Tehran Stock Exchange". *Int. Journal of Engineering Research and Applications*. ISSN: 2248–9622, Vol. 4, Issue 6(Version 2). pp.106–117.

Indicator Reference. 2016. http://www.fmlabs.com/reference/default.htm, (Accessed 27 March 2017).

James G., Witten D., and Hastie T., Tibshirani R. 2015, "An Introduction to Statistical Learning with R". *Springer*. USA. pp.175–197, pp.340–350.

Joshi K., Bharathi H.N., and Rao J. 2013., "Stock Trend Prediction Using News Sentiment Analysis", https://arxiv.org/ftp/arxiv/papers/1607/1607.01958.pdf (Accessed: 13–Apr–2017).

Keim D. 2006. "Financial Market Anomalie". *http://finance.wharton.upenn.edu/~keim/research/NewPalgraveAnomalies(May302006).pdf*.(Accessed: 14–Apr–2017).

Kuhn M., Johnson K. 2013. "Applied Predictive Modelling". *Springer*. USA. pp.69–73.

Lewis N.D. 2017, *Neural Networks for Time Series Forecasting with R: An* Intuitive Step *by* Step. Blueprint for Beginners. Kindle Edition. pp.141–179.

Niederhoffer V., and Osborne M. F. M. 1966. "Market making and reversal on the stock exchange". *Journal of the American Statistical Association*. 61(316). pp.897– 916.

Nikolaev N. (n.d.), "Multilayer Perceptrons (Continuation)". *http://homepages.gold.ac.uk/nikolaev/311bpr.htm.*(Accessed: 04–April–2017).

Maciel L.S., and Ballini R. n.d. "Design a neural network for time series financial
Forecasting: accuracy and robustness analysis".
https://www.academia.edu/4472250/DESIGN_A_NEURAL_NETWORK_FOR_TIME_SERIES_FINANCIAL_FORECASTING_ACCURACY_AND_ROBUSTNESS_ANALISYS (Accessed: 13–Jul–2017).

References (ctn'd)

Malkiel B. G. 1973. "A Random Walk Down Wall Street". https://www.academia.edu/10850809/A_Random_Walk_Down_Wall_Street_The_Time–Tested_Strategy_for_Successful_Investing.(Accessed: 13–Apr–2017)

Meesad P., and Li J. 2014. "Stock trend prediction relying on text mining and sentiment analysis with tweets". *In: 2014 Fourth World Congress on Information and Communication Technologies (WICT)*. pp. 257–262.

Murphy J. 1986. *Technical Analysis of the Financial Markets: A Comprehensive Guide to Trading Methods and Applications*, New York Institute of Finance, pp.225–262.

Olaniyan R., Stamate D., Lahcen O, and Logofatu D. 2015. "Sentiment and stock market volatility predictive modelling – A hybrid approach", In: Eric Gaussier; Longbing Cao; Patrick Gallinari; James Kwok; Gabriela Pasi and Osmar Zaiane, eds. *Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on.* Paris: IEEE, pp. 1–10.

Olaniyan R., Stamate D., and Logofatu D. 2015. "Social web–based anxiety index's predictive information on S&P 500 revisited", *Proceedings of the 3rd Intl. Symposium on Statistical Learning and Data Sciences*.

Parikh V. and Shah P. 2015. "Stock Prediction and Automated Trading System", *Computer Science & Electronics Journals*. Vol– 6, Issue–1 Sep. pp.104–111.

Rechenthin M.D. 2014. "Machine–learning classification techniques for the analysis and prediction of high–frequency stock direction". http://ir.uiowa.edu/cgi/viewcontent.cgi?article=5248&context=etd. (Accessed: 14–Apr–2017).

Sabri N.R. 2008. "The impact of trading volume on stock price volatility in the Arab economy", Journal of Derivatives & Hedge Funds. Volume 14, Issue 3–4. pp 285–298

Schoutens L. 2003. *Levy Processes in Finance*, West Sussex, England, p3, pp.27–28.

Schumaker R. P. and Chen H. 2009. "Textual analysis of stock market prediction using breaking financial news:The AZFinText system", *ACMTrans.Inform.Syst.27*, 2, Article12.

Senyurt G., and Subasi A. (n.d.), "Stock market movement direction prediction using tree algorithms" http://eprints.ibu.edu.ba/1187/1/41.%20Stock%20market%20movement%20direction%20prediction%20using%20tree%20algorithms.pdf, (Accessed: 14–Apr–2017).

Suresh A.S. 2013. "A Study on Fundamental and Technical Analysis", *International Journal of Marketing, Financial Services & Management Research*, Vol.2, No. 5.

Torgot L. 2011. *Data Mining with R: Learning with Case Studies*. Chapman & Hall/CRC. Florida. chp3. pp95–163.

Vaiz J.S., and Ramaswami M. 2016. "A Study on Technical Indicators in Stock Price Movement Prediction Using Decision Tree Algorithms", *American Journal of Engineering Research – (AJER)*. Volume 5, Issue 12. pp–207–212.

References (ctn'd)

Wang J., Wang J., Fang W., and Niu H. 2016, "Financial Time Series Prediction Using Elman Recurrent Random Neural Networks", *Computational Intelligence and Neuroscience.* Article ID 4742515.

Wiersma Y, Huettmann F., Drew A.C. 2011. "Predictive Species and Habitat Modeling in Landscape Ecology". https://books.google.co.uk/books?id=1V5gupaI5_IC&pg=PA147&lpg=PA147&dq=can+the+auc+and+kappa+disagree?&source=bl&ots=Pav9xYH5pT&sig=9cCUcQ5LLCAiOSkTqiw5BiUQ2qI&hl=en&sa=X&ved=0ahUKEwiPq67CrtHVAhXDCMAKHQ7MDt0Q6AEIKDAA#v=onepage&q=can%20the%20auc%20and%20kappa%20disagree%3F&f=false (Accessed 12 Aug 2017).

Q&A