STOCK PRICE INDEX MOVEMENT FORECASTING USING MACHINE LEARNING

Juliana Negrini de Araujo

FINANCIAL MARKET ANALYSIS

- Stock prices prediction is challenging
- Highly influenced by external factors: economic, political and market expectation
- Time-series model
- Forecast next day stock price direction
- Facebook, Apple and Google stocks analysed

DATASET

Classification problem

Data extracted from Yahoo! Finance

Initial features: Close, High, Low, Volume

Data of previous day predicts the market behaviour of tomorrow

$$y_t^i = \begin{cases} 1, if\ C_{(t)} > C_{(t-1)} \\ 0, if\ C_{(t)} \leq C_{(t-1)} \end{cases}$$

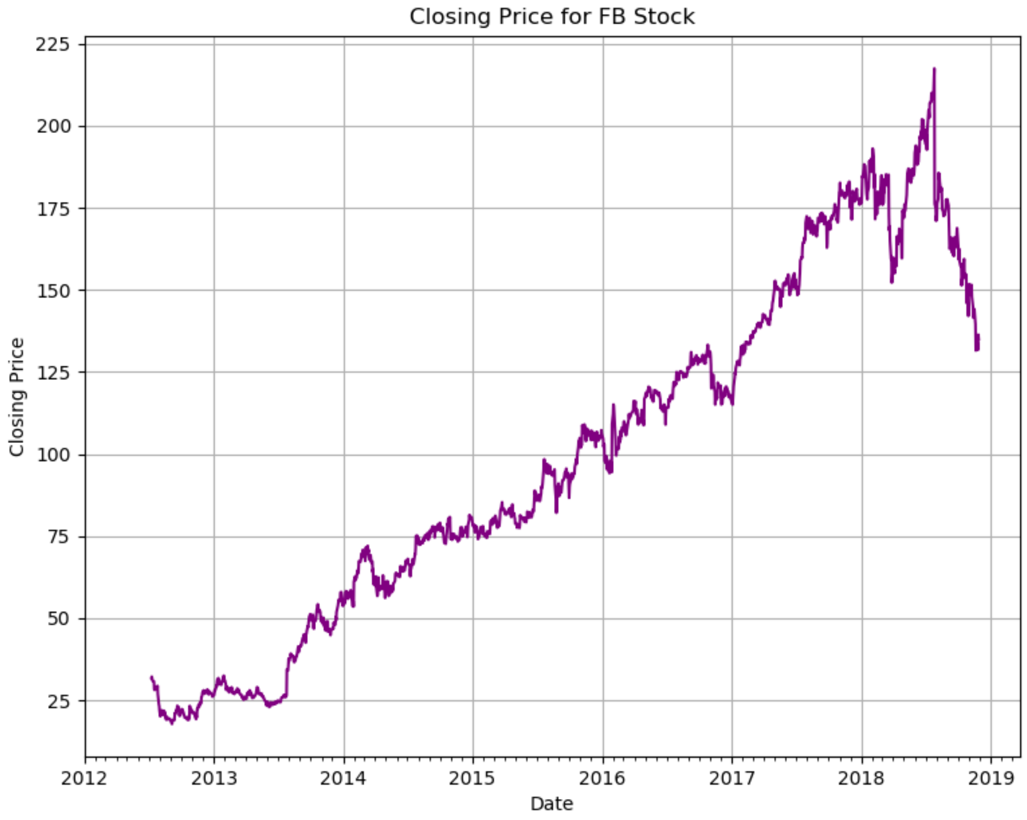
Date	C(t-1)	L(t-1)	H(t-1)	Vol (t-1)	Ct	Class
23/11/2018	176.78	176.55	180.27	31124200	172.29	0
26/11/2018	172.29	172.10	176.60	23624000	174.62	1
27/11/2018	174.62	170.26	174.95	44738600	174.24	0
28/11/2018	174.24	170.88	174.77	41387400	180.94	1

Previous day data

Closing price of current day

*Column removed from dataset prior to training / testing

DATASET - OVERVIEW



- Continuous upward trend;
- From 2018 and onwards closing price oscillates and downward trend initiates;
- ~4 years of data available;
- No data imbalance.

Stock	Data Range	Number of initial samples	% of class 1 samples
Facebook (FB)	18/05/2012 – 28/11/2018	1644	52.4%

DATASET

Preliminary tests could not achieve accuracy higher than **50%** with initial features (Close, High, Low, Volume);

Improve accuracy by adding new features:

- *Lag* features - Closing price of previous days as additional features
- Technical Indicators
- Global Market Index

DATASET

Technical Indicators added as additional features :

- Average
- Stochastic
- Momentum
- Overbought / Oversold

Moving m-day Average (MA)	$\frac{C_{(t-1)} + C_{(t-2)} + \cdots + C_{(t-m)}}{m}$
Weighted Moving m-day Average (WMA)	$\frac{(m)C_{(t-1)} + (m-1)C_{(t-2)} + \cdots + C_{(t-m)}}{m + (m-1) + \cdots + 1}$
Momentum (M)	$C_{(t-1)} + C_{(t-m)}$
Stochastic Oscillator (SO)	$\frac{C_{(t-1)} - LL_{(t-(m-1))}}{HH_{(t-(m-1))} - LL_{(t-(m-1))}} x100$
Moving Stochastic Oscillator (SSO)	$\frac{1}{m}\sum_{i=t-m+1}^{t}(SO_{(t-1)})$
Exponential Moving Average (EMA)	$\alpha C_{(t-1)} + EMA(k)_{(t-1)}$
Moving Average Convergence Divergence (MACD)	$MACD(k)_{(t-1)}$ $+\frac{2}{k+1}[(EMA(12)_{(t-1)} - EMA(26)_{(t-1)}) - MACD(k)_{(t-1)}]$
Relative Strength Index (RSI)	$100 - \frac{100}{1 + RS}$
Commodity Channel Index (CCI)	$\frac{M_{(t-1)} - SM_{(t-1)}}{0.015D_{(t-1)}}$
Accumulation / Distribution Oscillator (ADO)	$\frac{(C_{(t-1)} - L_{(t-1)}) - (H_{(t-1)} - C_{(t-1)})}{H_{(t-1)} - L_{(t-1)}}$

DATASET

Technical Indicators as trend deterministic data:

- Average
- Stochastic
- Momentum
- Overbought/Oversold

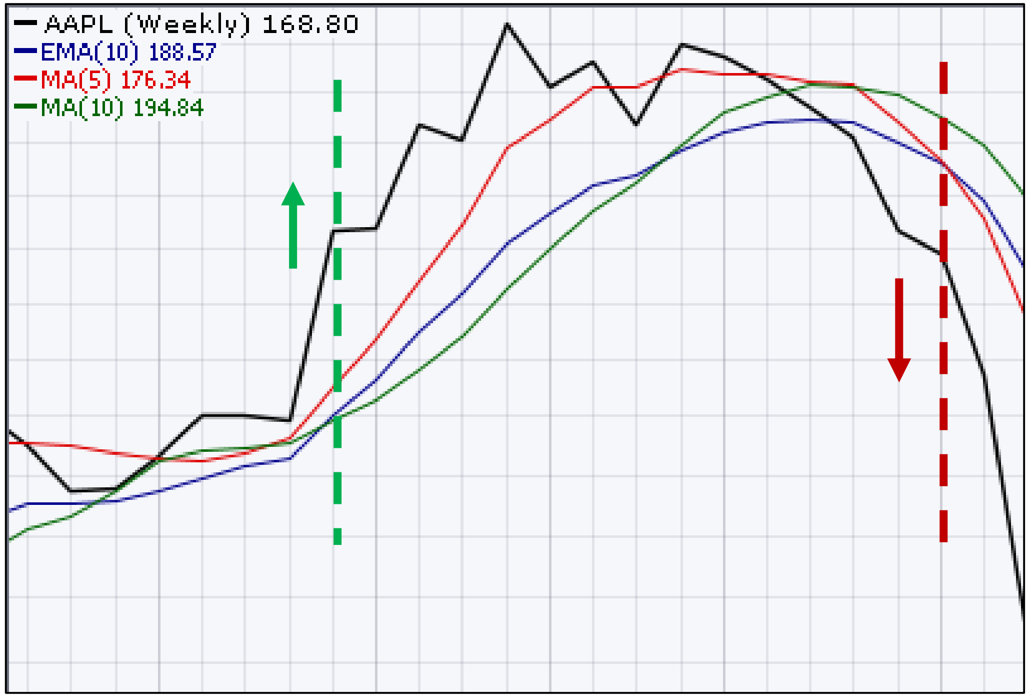
Example:
If RSI is above 70, output 0, as it indicates a overbought. Values below 30 output is one, indicating oversold. For values within this range it compares to previous day and outputs 1 if value has increased;

RSI	CCI
61.61	270.27
62.43	184.60
63.39	125.17
60.55	119.71
61.93	103.80
62.01	96.36
62.71	91.33
65.97	99.81
66.46	93.77
67.44	91.41
70.62	106.64
76.10	160.11
77.09	167.51
80.07	183.88

RSI >70 <30	CCI >200 <-200
1	0
1	0
1	0
0	0
1	0
1	0
1	0
1	1
1	0
1	0
0	1
0	1
0	1
0	1

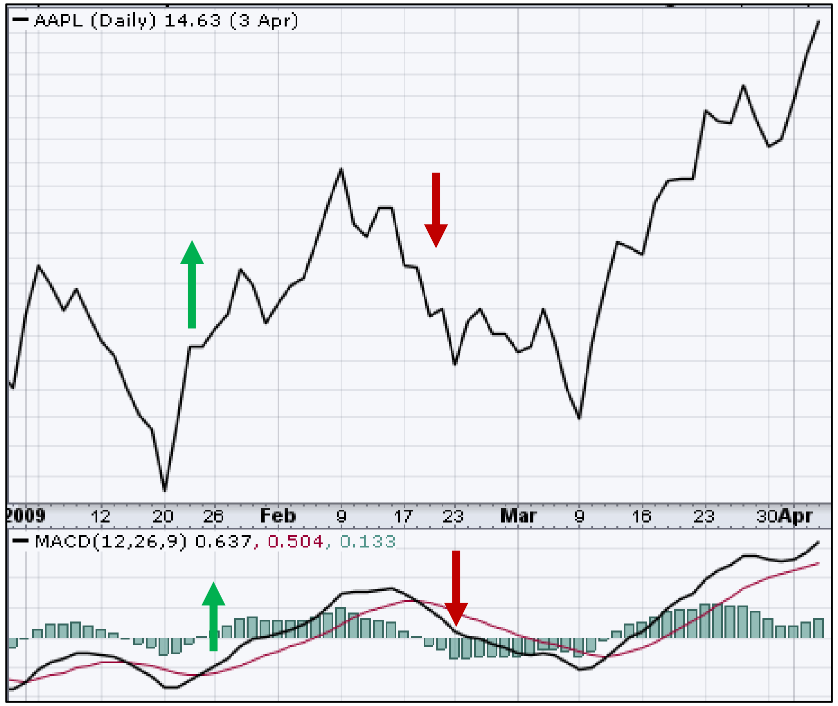
DATASET

Moving Averages
$$\begin{cases}1, if\ C_{(t)} > MA_{(t-1)}\\0, if\ C_{(t)} \leq MA_{(t-1)}\end{cases}$$



Example:
When the MA value is lower than closing price it can be interpreted as an indication that the stock value will go up;

MACD
$$\begin{cases}1, if\ MACD_{(t-1)} > MACD_{(t-2)}\\0, if\ MACD_{(t-2)} \leq MACD_{(t-1)}\end{cases}$$



Example:
MACD trend follows the stock trend. An upward MACD trend also indicates a rise in the stock price.

DATASET

Technical Indicators as trend deterministic data added as additional features:

- Average
- Stochastic
- Momentum
- Overbought/Oversold

Moving m-day Average (MA)	$\begin{cases} 1, if\ C_{(t)} > MA_{(t-1)} \\ 0, if\ C_{(t)} \leq MA_{(t-1)} \end{cases}$
Weighted Moving m-day Average (WMA)	$\begin{cases} 1, if\ C_{(t)} > WMA_{(t-1)} \\ 0, if\ C_{(t)} \leq WMA_{(t-1)} \end{cases}$
Exponential Moving Average (EMA)	$\begin{cases} 1, if\ C_{(t)} > EMA_{(t-1)} \\ 0, if\ C_{(t)} \leq EMA_{(t-1)} \end{cases}$
Momentum (M)	$\begin{cases} 1, if\ M_{(t-1)} > 0 \\ 0, if\ M_{(t-1)} \leq 0 \end{cases}$
Stochastic Oscillator (SO)	$\begin{cases} 1, if\ SO_{(t-1)} > SO_{(t-2)} \\ 0, if\ SO_{(t-1)} \leq SO_{(t-2)} \end{cases}$
Moving Stochastic Oscillator (SSO)	$\begin{cases} 1, if\ SSO_{(t-1)} > SSO_{(t-2)} \\ 0, if\ SSO_{(t-1)} \leq SSO_{(t-2)} \end{cases}$
Moving Average Convergence Divergence (MACD)	$\begin{cases} 1, if\ MACD_{(t-1)} > MACD_{(t-2)} \\ 0, ifMACD_{(t-1)} \leq MACD_{(t-2)} \end{cases}$
Accumulation / Distribution Oscillator (ADO)	$\begin{cases} 1, if\ ADO_{(t-1)} > ADO_{(t-2)} \\ 0, if\ ADO_{(t-1)} \leq ADO_{(t-2)} \end{cases}$
Relative Strength Index (RSI)	$\begin{cases} 1, if\ RSI_{(t-1)} < 30 \\ 0, if\ RSI_{(t-1)} > 70 \end{cases}$
Commodity Channel Index (CCI)	$\begin{cases} 1, if\ RSI_{(t-1)} < -200 \\ 0, if\ RSI_{(t-1)} > 200 \end{cases}$

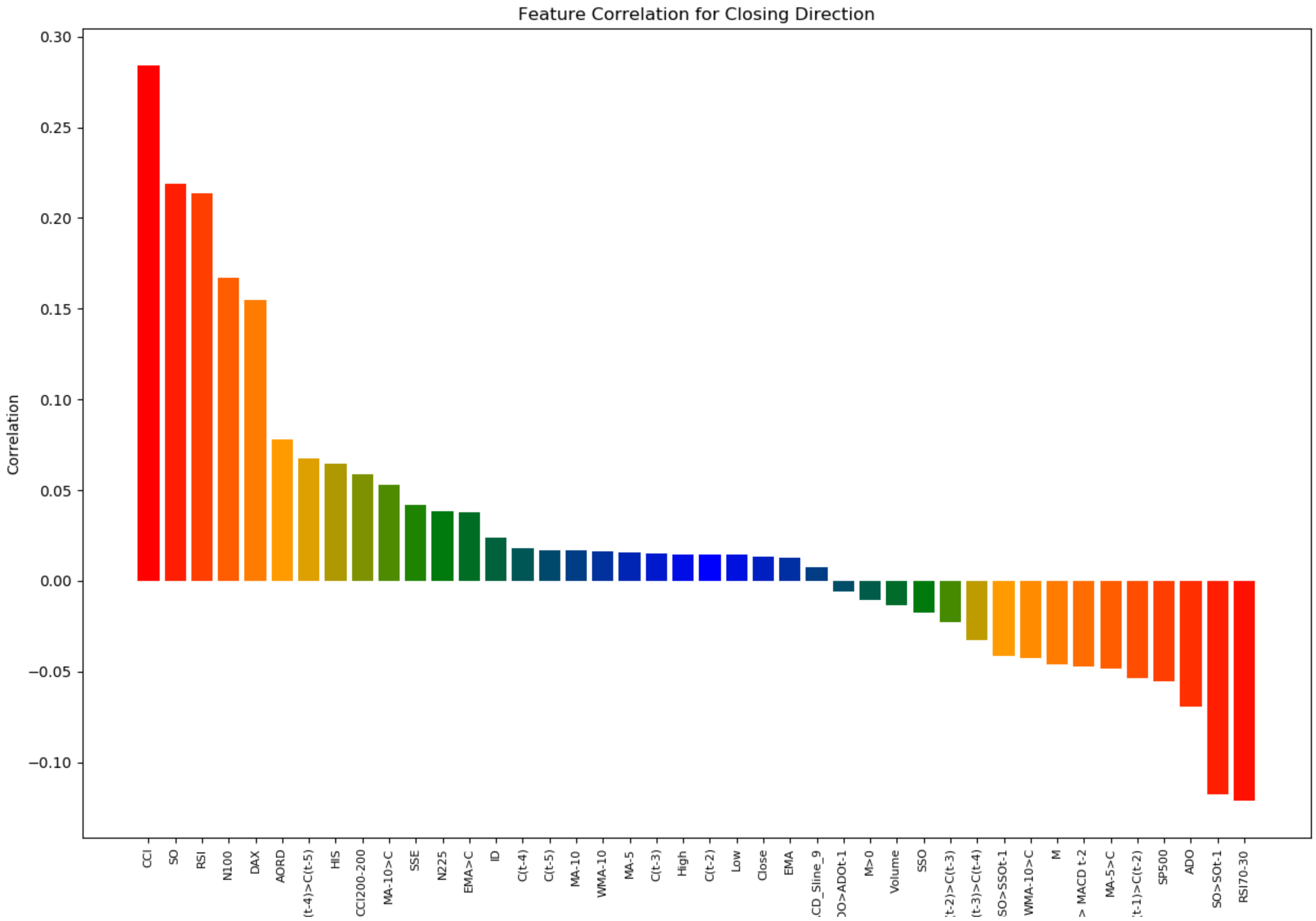
DATASET

Global Indexes - As part of a globalised economy, foreign markets daily performance can influence the behaviour of the selected American stocks.

Index	Description
Nikkei 225	Trades on Tokyo Stock Exchange and contains main 225 Japanese companies.
Hang Seng	Hong Kong market index containing major local companies.
All Ordinaries	Index share containing 500 Australian companies.
Euronext 100	Index comprising the 100 largest Euronext stocks. Contains companies from France, Netherlands, Belgium, Portugal and Luxembourg.
SSE Composite Index	Largest Chinese stock exchange, this index represents all shares traded in Shanghai Stock Exchange.
DAX	Trades on Frankfurt Stock Exchange and comprises 30 major German businesses.

- Global Indexes and Technical Indicators are within the top correlated features for this dataset;

- Low correlation value of Close, Low, High (in blue) could be an explanation of why the initial model had such low performance;

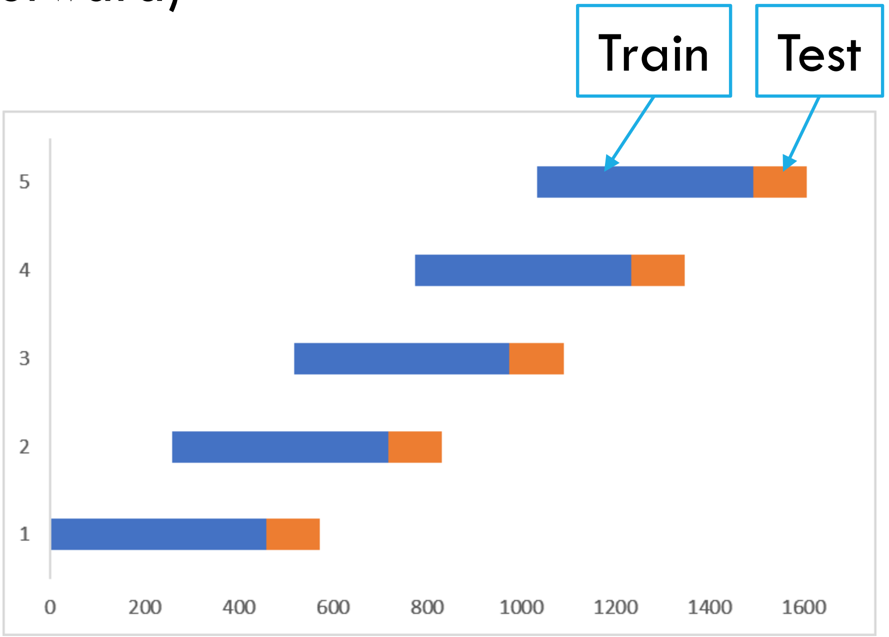


TRAINING THE MODEL

SPLITTING DATASET AND PARAMETER TUNING

- K-Cross validation is not applicable for time series data problems
- Best performance achieved with Sliding Window (walk-forward)

- Splitting the dataset into N splits (N windows-N-training sets, N testing sets)
- Training each split window for set of parameters
- Finding best fitted parameters for each window (eg. Gamma, C for SVM model)
- Each window will have different set of parameters giving better accuracy
- Fitting best parameters from previous step for all windows
- This is to check which parameter set is giving better overall results
- Averaging the accuracy of all window to chose the best parameters



TRAINING THE MODEL

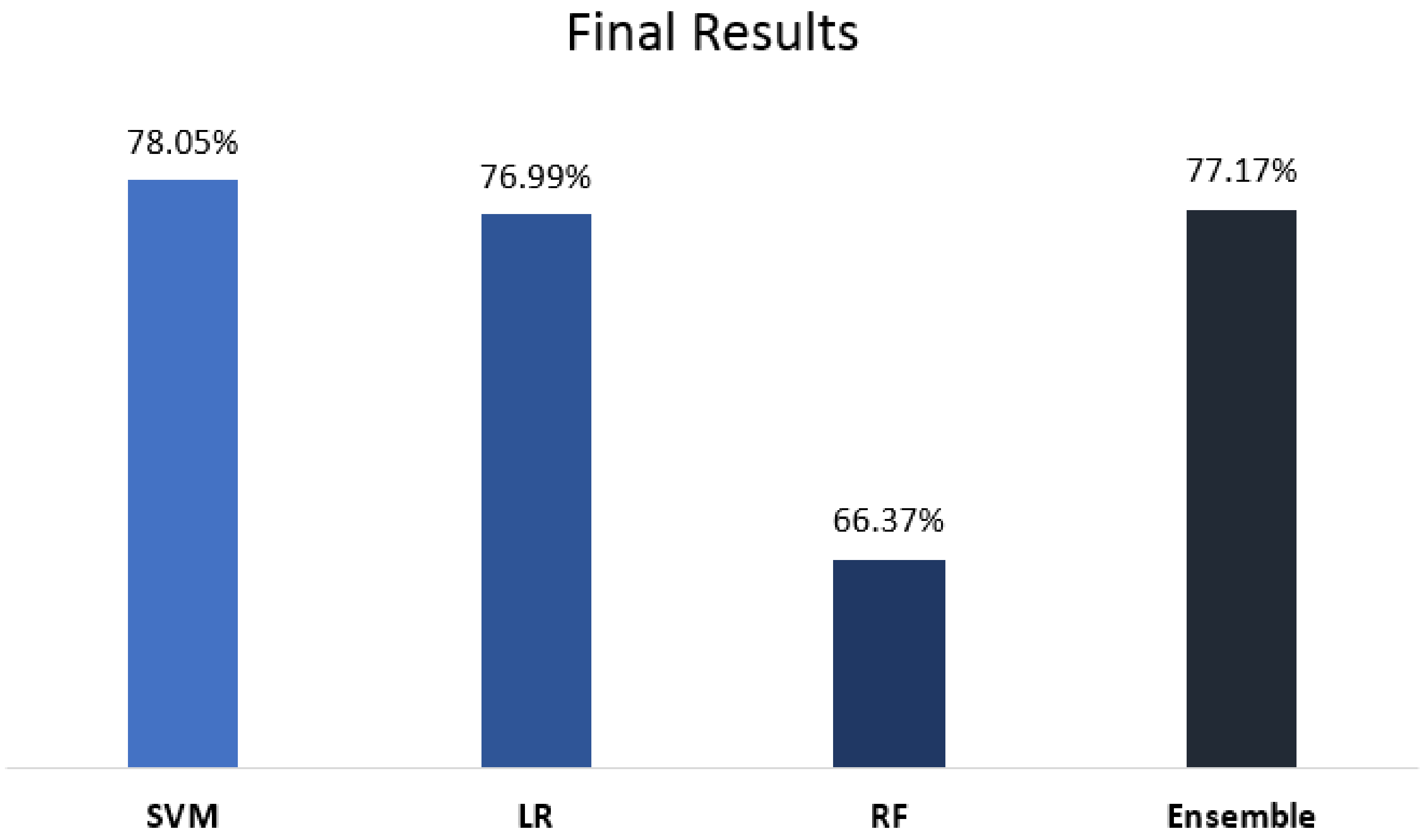
Selection of 3 ML classifier algorithms

1. Random Forest (criterion= 'gini', max_depth= 2, n_estimators= 10)
2. Logistic Regression (solver = 'liblinear', penalty = 'l1',C = reg_C,max_iter = 5000)
3. Support Vector Machines (kernel= 'rbf', C= 60, gamma= 0.001)
4. Ensemble model of SVM and Logistic Regression (Random Forest excluded due to poor performance)

The final parameters were selected as the ones that provided the best accuracy considering the average accuracy from all cross validation windows.

RESULTS

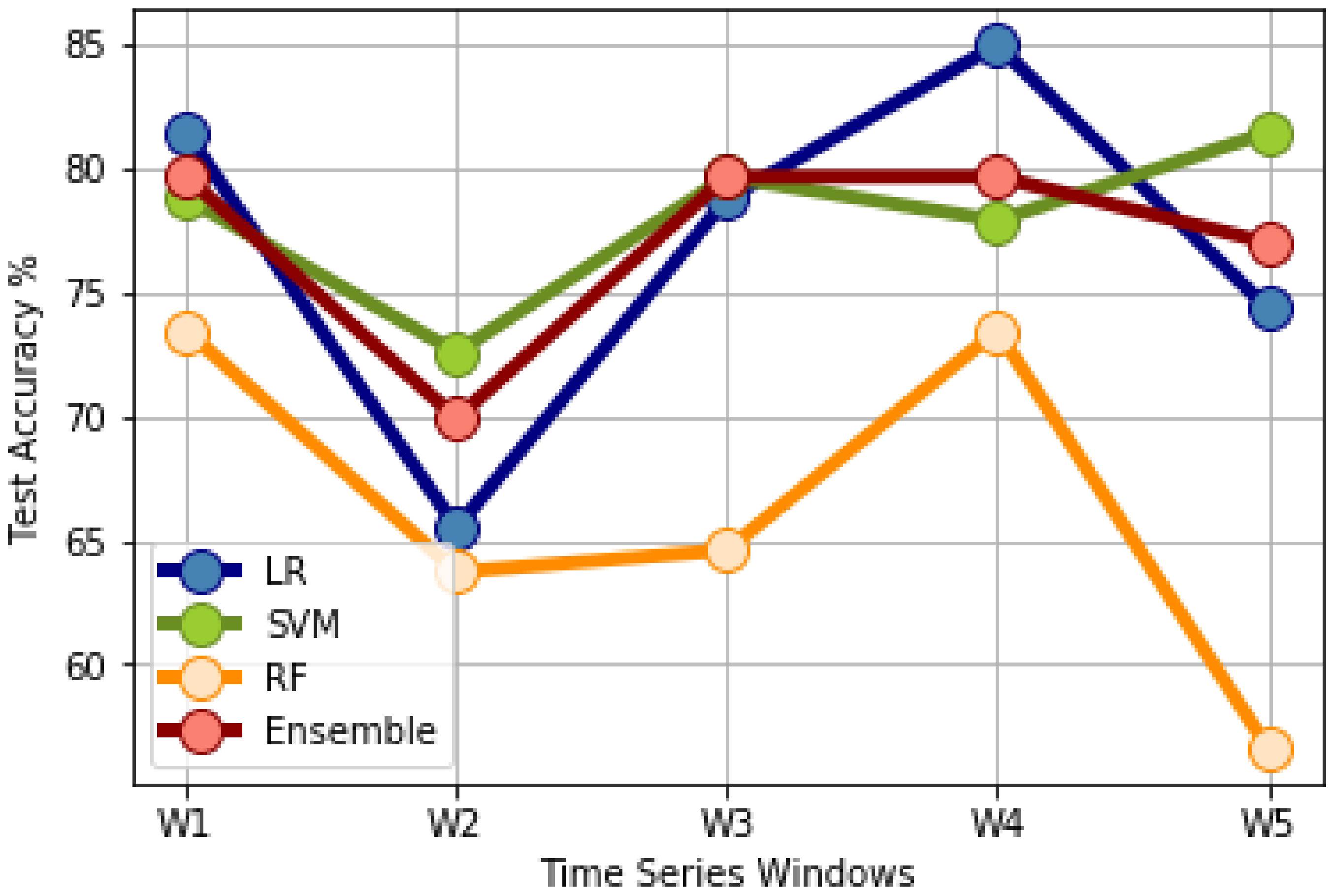
Accuracy = (TP+TN)/(TP+TN+FP+FN)



RESULTS

- SVM has the lower standard deviation and higher mean accuracy;
- W2 was the most challenging period to predict, as most models presented their lower accuracy on this timeframe;
- From the graph the red line is the middle ground between the SVM and LR results. However, the ensemble model presented lower accuracy overall, even though an effort was made to weight the contribution of both models.

Accuracy results across the five windows



DISCUSSION AND CONCLUSION

- The results have shown that SVM outperforms logistic regression and random forest algorithms for stock movement prediction. The inherent capability of SVM to avoid overfitting contributed to this conclusion.
- There was no advantage on using the Ensemble model, as it was not succesfull to merge the positive characteristics of both models (LR and SVM).
- During experimental testing, random forest showed a high tendency to overfit to training data achieving contradictory results between training and test accuracy.
- Essential contributions from literature allowed the construction of the dataset that allowed the models to forecast with similar accuracies of current published papers.
- As it is the case for market traders, the usage of technical indicators and global indexes have shown to be a powerful strategy to support forecast decisions.
- Sliding window provided better when compared to results as it considers more recent data as input for each window.

FUTURE WORK

- Optimise the sliding window training/test set sizes;
- Usage of ANN or the specialised LSTM for time series;
- Analyse other indicators available to improve accuracy.