PREDICTING STOCK MARKET RETURNS

Divya Naidu

Objective

IMPORTANCE

- Major Investment Activity
- Ability to forecast which way the market is going to move

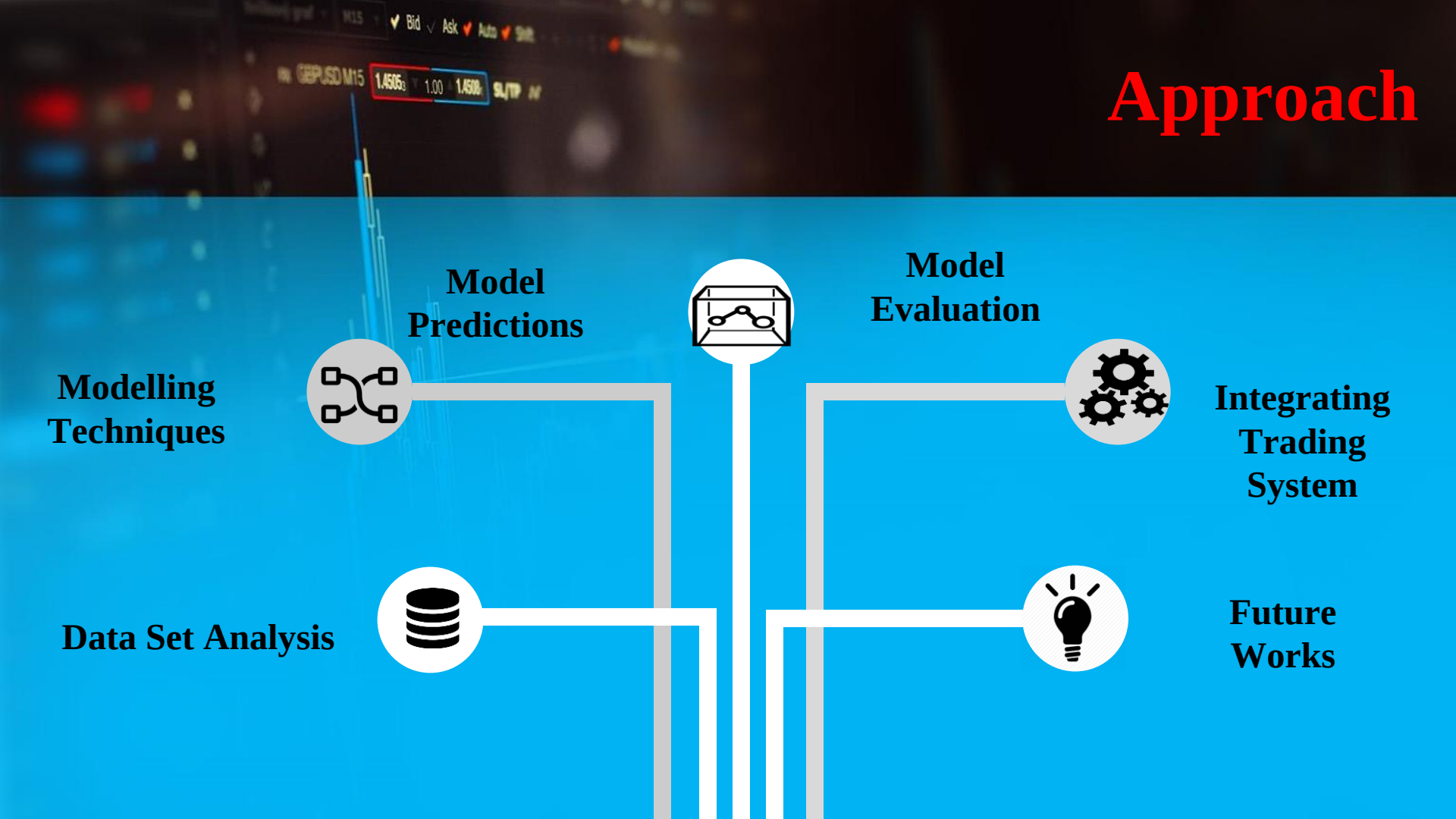
METHODOLOGY

- Human inspection of data
- Data mining models

GOAL

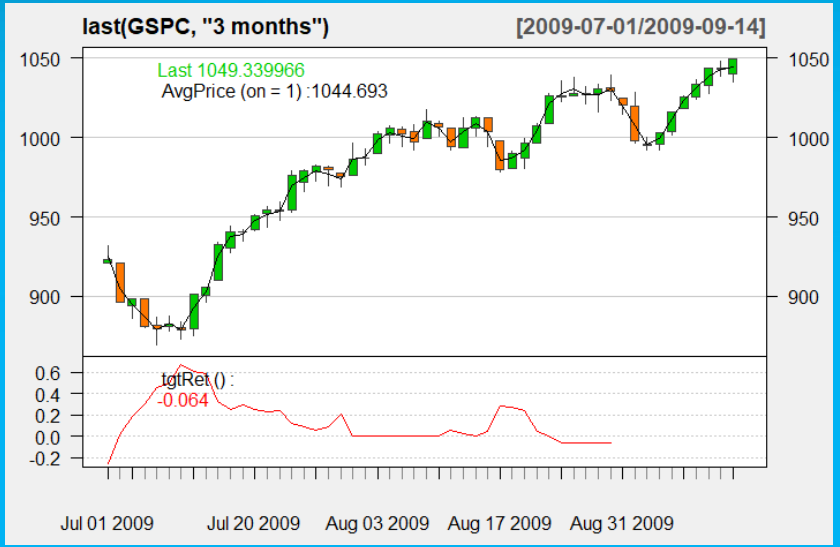
- Build a stock trading system for S&P 500 market index using data mining techniques on daily stock quotes data
- Good Prediction of the overall tendency of the prices.

Approach


Model Predictions
Model Evaluation
Modelling Techniques
Integrating Trading System
Data Set Analysis
Future Works

Data Set Analysis

1. Multivariate Financial time series data
2. Date,Open,High,Low.Close,Volume,AdjClose
3. Variable ‘T’ Indicates the tendency of the market
4. Threshold (p%) for T indicator = 2.5%
5. p% > 2.5% = ’BUY’
   p% < 2.5% = ‘SELL’



Indicators

The indicators usually try to capture some properties of the prices series such as if they are varying too much, or following some specific trend, etc.

Approach

- Feature Filters
- Feature wrappers.

Feature Filters: Use some statistical properties of the features to select the final set of f features.

Feature wrappers:- The wrapper approaches include the tool and an iterative search process in the selection process.

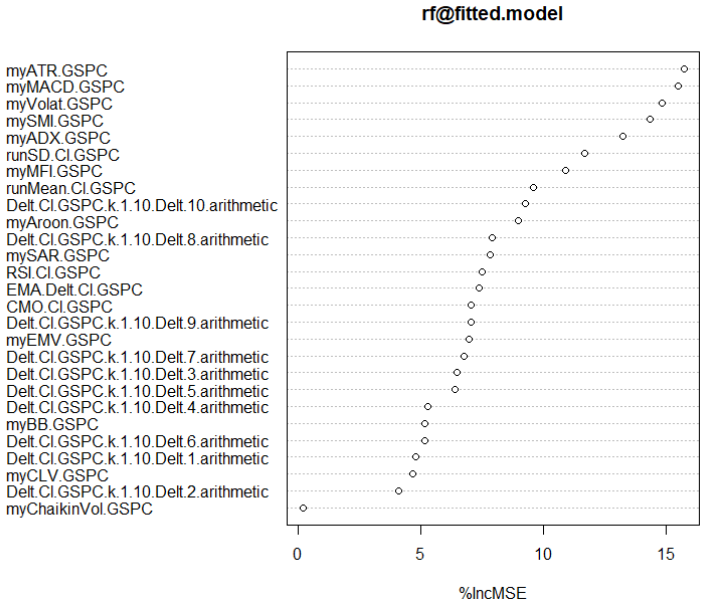
22 variables (myATR, mySMI etc) ⟹ 7 top variables (score > 10)

We will split the available data into two separate sets:-

- One used for constructing the trading system (30 years)
- Final Evaluation set (9 years)

Top Indicators

- At first a quantmod object that contains the specification of a certain abstract model is created
- We used Random forest to estimate the importance of variables involved in the prediction tasks.
- We have used 10 as the threshold on the importance scores




```
[1] "Delt.Cl.GSPC.k.1.10.Delt.1.arithmetic"
[2] "myATR.GSPC"
[3] "myADX.GSPC"
[4] "myEMV.GSPC"
[5] "myVolat.GSPC"
[6] "myMACD.GSPC"
[7] "mySAR.GSPC"
[8] "runMean.Cl.GSPC"
```

Prediction Tasks

- There are two paths to obtain prediction for the correct trading signal
- The first part uses the T value as the target variable and tries to obtain models that forecast this value using the predictors information.

$$signal = \begin{cases} sell & \text{if } T < -0.1 \\ hold & \text{if } -0.1 \leq T \leq 0.1 \\ buy & \text{if } T > 0.1 \end{cases}$$

- In the second alternative prediction task we consider of predicting the signals directly.
- Regression tasks have a numeric target variable (e.g., our T indicator), while classification tasks use a nominal target variable, that is, with a finite set of possible values.

Training Data and Evaluation Criteria

- Event-based prediction tasks are usually evaluated by the precision and recall metrics.
- Precision can be informally defined as the proportion of event signals produced by the models that are correct.
- Recall is defined as the proportion of events occurring in the domain that is signaled as such by the models.
- Regime shifts are large abrupt, persistent changes in the structure and function of a system.

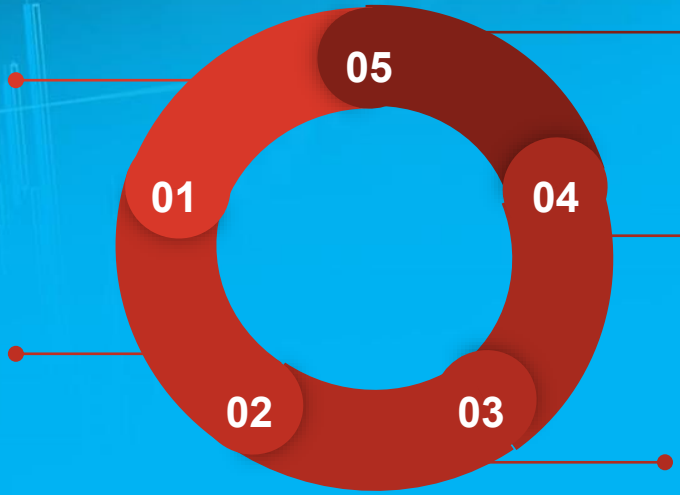
A Confusion Matrix for the Prediction of Trading Signals

		Predictions			
		sell	hold	buy	
True Values	sell	$n_{s,s}$	$n_{s,h}$	$n_{s,b}$	$N_{s,.}$
	hold	$n_{h,s}$	$n_{h,h}$	$n_{h,b}$	$N_{h,.}$
	buy	$n_{b,s}$	$n_{b,h}$	$n_{b,b}$	$N_{b,.}$
		$N_{.,s}$	$N_{.,h}$	$N_{.,b}$	N

Why ANN?

Ability to learn and model non-linear and complex relationships

ANN can generalize that is it can infer unseen relationships or unseen data as well.

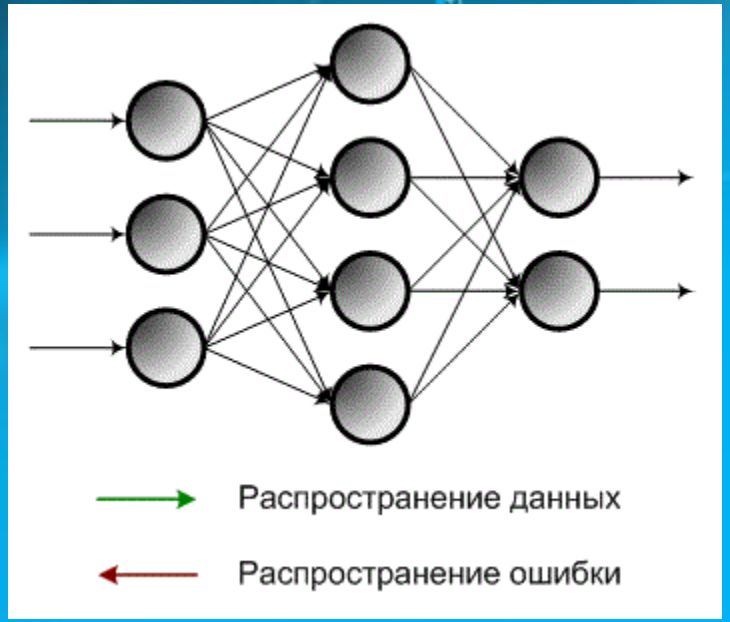


ANN does not impose any restrictions on the input variables.

It has the ability to take lots of inputs and process them to infer hidden as well as complex non-linear relationships.

ANN applied in the right way can provide robust alternative, given its ability to model and extract features and relationships.

Artificial Neural Network



ANNs are formed by a set of computing units (the neurons) linked to each other.

Constructing an artificial neural network consists of establishing an architecture for the network and then using an algorithm to find the weights of the connections between the neurons.

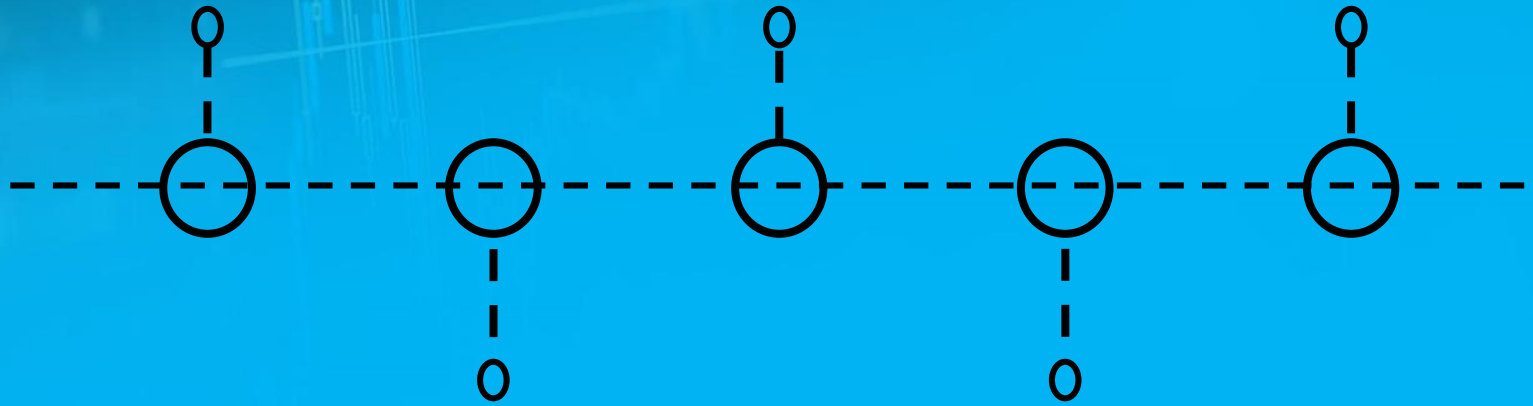
ANNs are known to be sensitive to different scales of the variables used in
a prediction problem.

Why SVM ?

SVM models have generalization in practice, the risk of overfitting is less

It scales relatively well to high dimensional data

Training a SVM involves optimization of a convex function with linear constraint



The constructed model has an explicit dependence only on the support vectors, which reduces the computational cost.

With an appropriate kernel function, any complex problem can be solved

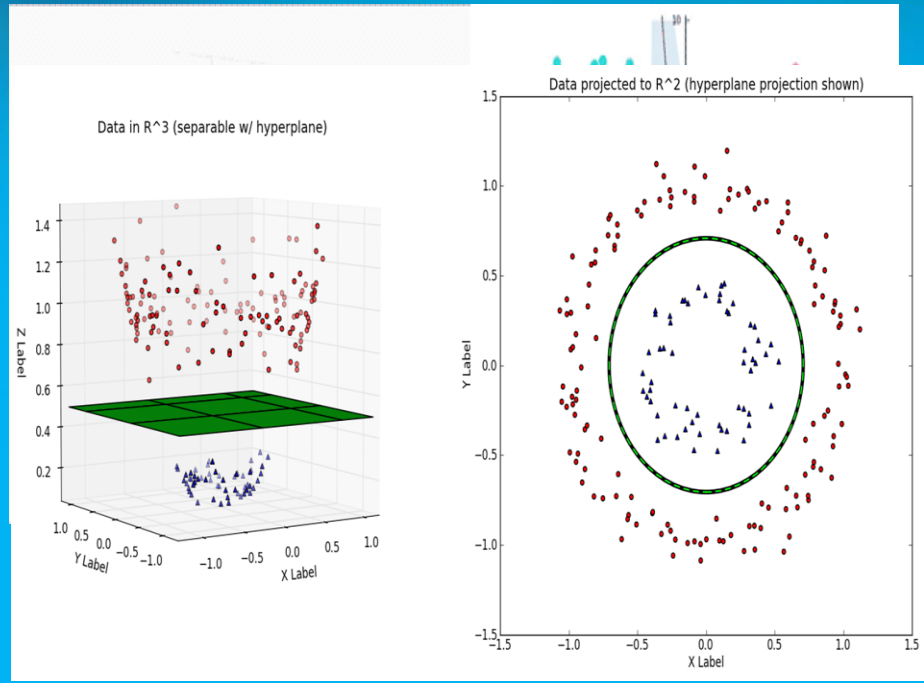
Support Vector Machine

The basic idea behind SVMs is that of mapping the original non linear data into a new, high-dimensional space, where it is possible to apply linear models to obtain a separating hyperplane.

Soft margin methods allow for a small proportion of cases to be on the "wrong" side of the margin, each of these leading to a certain "cost"

SVM REGRESSION - the SVM model achieves a considerably better score than the ANN in terms of precision, although with a much lower recall.

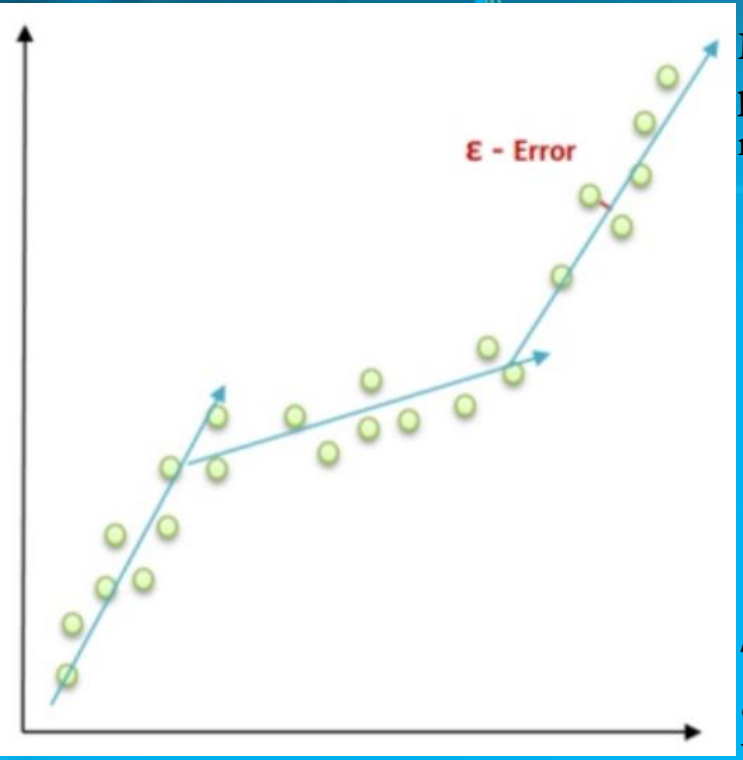
SVM CLASSIFICATION - The results of this SVM are not as interesting as the SVM obtained with the regression data



Why MARS?

- Works well with a large number of predictor variables
- Automatically detects interactions between variables
- It is an efficient and fast algorithm, despite its complexity
- Robust to outliers

Multivariate Adaptive Regression Splines



Multivariate Adaptive Regression Splines (MARS) is a non-parametric regression method that builds multiple linear regression models across the range of predictor values.

We fit linear line on only a small part of line(spline) .We then create a pivot point and then second linear line is added connecting to the first line

The basis functions can take several forms, from simple constants to functions modeling the interaction between two or more variables.

A hinge function is the point where the linear regression models line is shifted into a different linear regression line

The results are comparable to the ones obtained with SVMs for classification, with precision scores around 30%, although with lower recall

Prediction Into Action

Trading Strategies

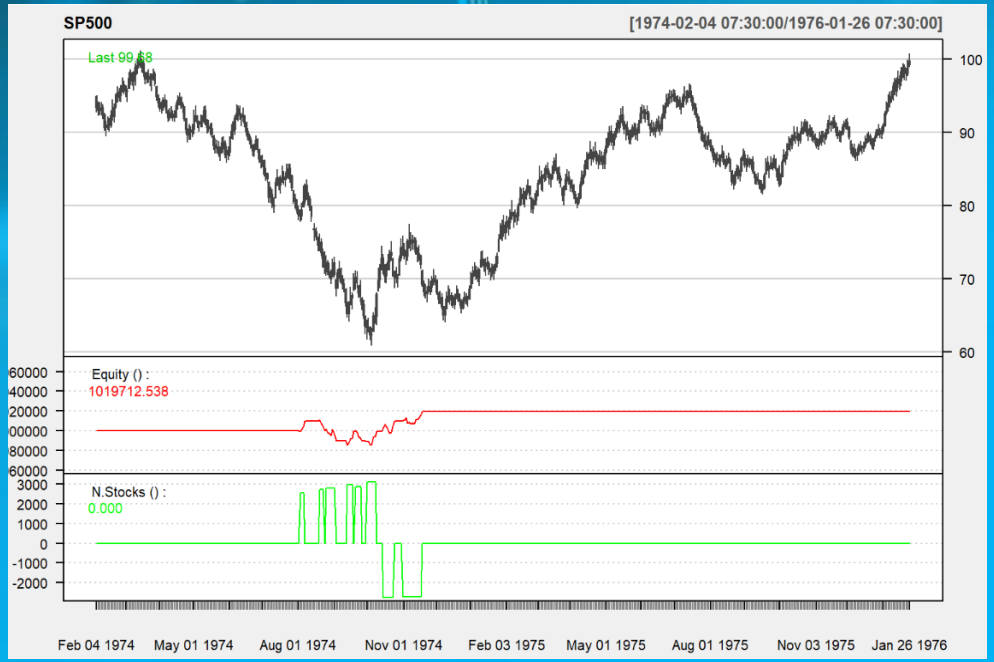
- All decisions will be taken at the end of the day.
- If at the end of the day we have a prediction for a low value of T
  - When this order is carried out in the future at a price p we immediately post two extra orders:
    - A buy limit order with a limit price of p - t% (t% is our target profit) with a deadline of 10 days
    - A buy stop order with a limit price of p + l% (l% is our maximum accepted loss)
- If the models forecast a high value of T
  - When this order is carried out in the future at a price p we immediately post two extra orders:
    - A sell limit order with a limit price of p + t% (t% is our target profit) with a deadline of 10 days
    - A sell stop order with a limit price of p - l% (l% is our maximum accepted loss)

A second possible strategy, which is similar to the 1st strategy but with two exceptions, is one where we will always open new positions even if we have already an opened position we will wait for ever for positions to reach the target profit.

Prediction Into Action

Trading Strategies

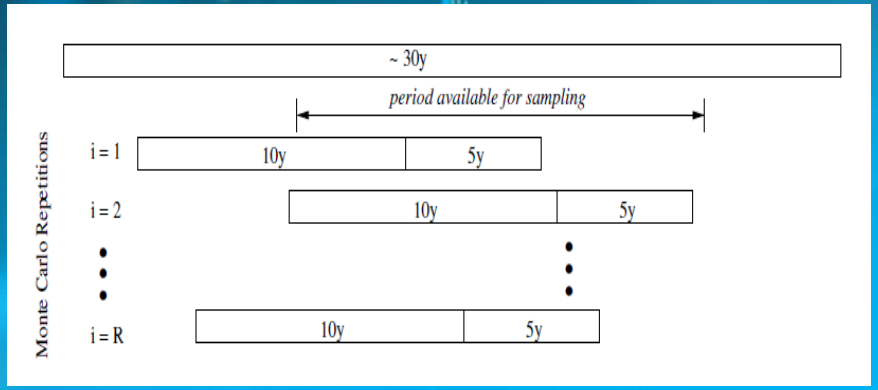
The results of trading using Policy 1 based on the signals of an SVM



tradingEvaluation() used to obtain a series of economic indicators of the performance

Using the same signals but with a different trading policy the return decreased from -1.04% to -7.89%

Model Evaluation



Our dataset includes around 30 years of daily quotes

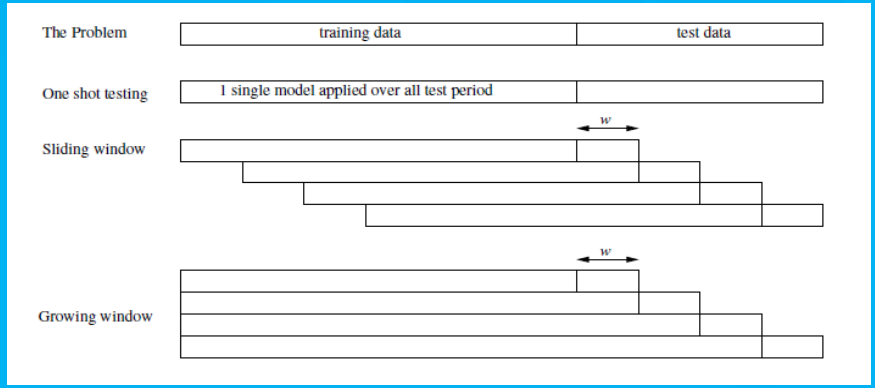
Train on 10 years of data | Test on 5 years of data

The function *trading.simulator()* will be used to obtain the training record

One Shot Testing: Single model on training data to develop the model and obtain predictions for the testing period

Sliding Window: This approach deletes the oldest data of the training set at the same time it incorporates the fresher observations

Growing Window: This approach adds the test data to the current training set, does not delete the oldest data.



Model Evaluation

We performed 20 iterations on 10 years of training data and 5 years of test data to implement Monte Carlo

We implemented different variants of of all possible combinations of the values for the parameters in the models

Each of these variants will then be run in three different model updating“modes”: single, sliding window, and growing window.

Moreover, we will try for the two latter modes two relearn steps: 60 and 120 days.

VARIANTS: 60 SVM | 120 MARS | 60 ANN

Result Analysis

- Minimal values constraints:
  (1) a reasonable number of average trades, say more than 20
  (2) an average return that should at least be greater than 0.5% (given the generally low scores of these systems)
  (3) and also a percentage of profitable trades higher than 40%.
- Out of the 240 trading variants compared only 3 satisfied the minimal values constraints.
- All the three use regression tasks and are based on Neural networks.
- According to the Wilcoxon significance test single.nnetR.v12 method has the highest average return with 95% confidence.

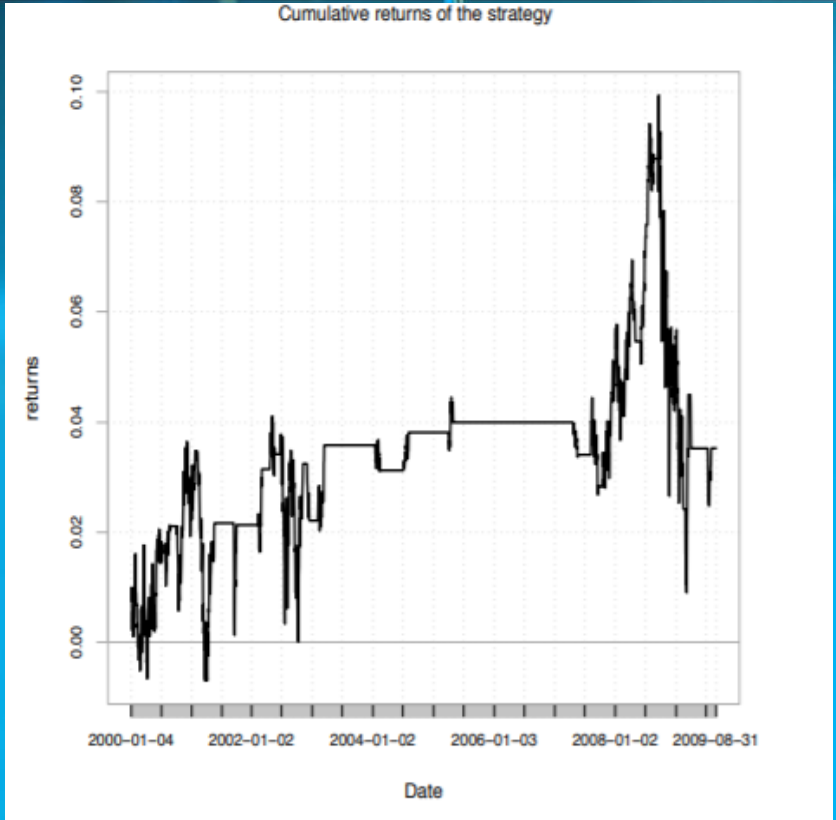
The Trading System

Evaluating the final test data

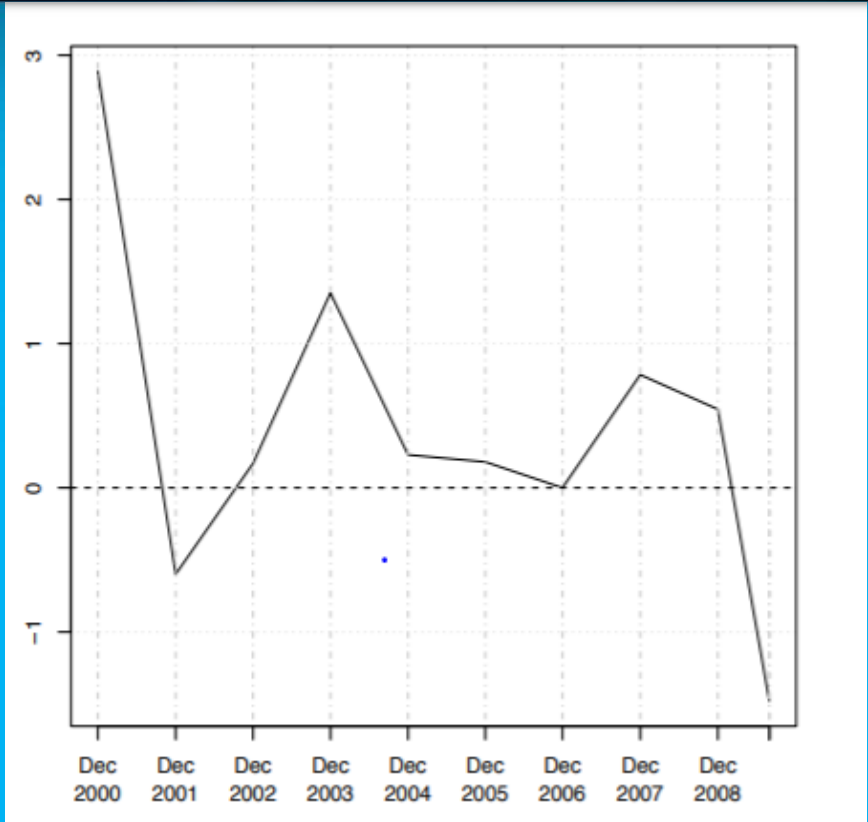
- The final evaluation on the 3 best models showed that the trading variant based on ANN regression and using the growing window schema was the best model.

- We then performed a detailed analysis of the model across the evaluation period.

The cumulative returns on the final evaluation period of the system



Yearly percentage returns of system



Online Trading System

- We intend to use the trading system developed by us in real time to trade on the market.
- The general algorithm of the system implementation is:
  - Read in the current state of the trader
  - Get all new data available
  - Check if it is necessary to re-learn the model
  - Obtain the predicted signal for today
  - With this signal, call the policy function to obtain the orders
  - Output the orders of today

Conclusion

- **It was observed that the trading variant model based on ANN regression using the growing window schema was observed to be the best model.**

- **Using this model we can predict the tendency of S & P 500's market index for our Online Trading System.**

Future Works

- **Predict the actual values of the market index.**
- **Make use of additional data sets.**
- **Implement more models based on neural networks**
- **Use logistic regression modelling.**