AL/SM/90/01 01.02.2022

Market Direction Prediction using Statistical Models & Machine learning Models

Introduction:

This study attempts to predict Stock Market Index Direction based on 8 Predictor variables.

Objective of the study is to compare the Predictive Power of the Statistical Model Vs Machine Learning Models, on Financial Time series data.

S&P 500 Index data for 5 years was taken and for each date, the daily return and the returns for previous days (Lag1…Lag5) were computed . The direction of the market (Up, Down) along with Trading volume were recorded.

The data set was preprocessed and was split in 2 sets.

a) Training data set – 1000 observations
b) Test data set- 250 observations

The Models were built on training data set and tested on test data set .

For each Model, a confusion matrix was created and misclassication error rate was computed so as compare the Models.

The Statistical Models chosen:

- Logistic Model
- LDA Model
- QDA Model
- Naïve Bayes Model

The Machine learning Models used:

- SVM
- Random Forest
- Neural network

Data Set:

```
data.frame':1250 obs. of  9 variables:
 $ Year    : num  2001 2001 2001 2001 2001 ...
 $ Lag1    : num  0.381 0.959 1.032 -0.623 0.614 ...
 $ Lag2    : num  -0.192 0.381 0.959 1.032 -0.623 ...
 $ Lag3    : num  -2.624 -0.192 0.381 0.959 1.032 ...
 $ Lag4    : num  -1.055 -2.624 -0.192 0.381 0.959 ...
 $ Lag5    : num  5.01 -1.055 -2.624 -0.192 0.381 ...
 $ Volume: num  1.19 1.3 1.41 1.28 1.21 ...
 $ Today   : num  0.959 1.032 -0.623 0.614 0.213 ...
 $ Direction: Factor w/ 2 levels "Down","Up": 2 2 1 2 2 2 1 2 2 2 ...
```

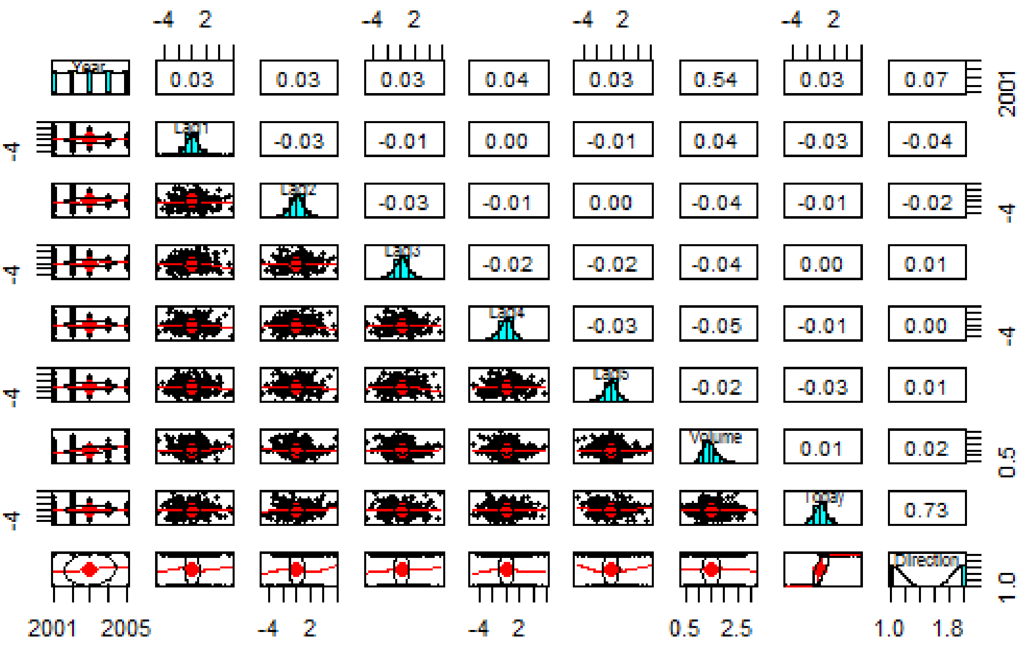
Summary(Data Set)

```
 Lag1                    Lag2                          Lag3
 Min.   :-4.922000       Min.   :-4.922000              Min.   :-4.922000
 1st Qu.:-0.639500       1st Qu.:-0.639500               1st Qu.:-0.640000
 Median : 0.039000        Median : 0.039000             Median : 0.038500
 Mean   : 0.003834       Mean   : 0.003919              Mean   : 0.001716
 3rd Qu.: 0.596750       3rd Qu.: 0.596750              3rd Qu.: 0.596750
 Max.   : 5.733000       Max.   : 5.733000              Max.   : 5.733000

 Lag4                    Lag5                   Volume          Today
 Min.   :-4.922000       Min.   :-4.92200       Min.   :0.3561    Min.   :-4.922000
 1st Qu.:-0.640000       1st Qu.:-0.64000      1st Qu.:1.2574     1st Qu.:-0.639500
 Median : 0.038500        Median : 0.03850      Median :1.4229     Median : 0.038500
 Mean   : 0.001636       Mean   : 0.00561       Mean   :1.4783     Mean   : 0.003138
 3rd Qu.: 0.596750       3rd Qu.: 0.59700       3rd Qu.:1.6417     3rd Qu.: 0.596750
 Max.   : 5.733000       Max.   : 5.73300       Max.   :3.1525     Max.   : 5.733000
```

Direction :Down:602 , Up :648

Distribution & Correlation:



Statistical Models Classification Performance

1) Logistic Model:

Confusion Matrix

	Actual	
Model	Down	Up
Down	2	1
Up	109	140

Misclassification rate: 0.4365079

2) Linear Discriminant Analyses Model:

Confusion Matrix

	Actual	
Model	Down	Up
Down	2	12
Up	108	130

Misclassification rate: 0.47619

3) Quadratic Discriminant Analyses Model:

Confusion Matrix

	Actual	
Model	Down	Up
Down	43	51
Up	68	90

Misclassification rate: 0.4722222

4) Naive Bayes Model:

Confusion Matrix

	Actual	
Model	Down	Up
Down	36	45
Up	75	96

Misclassification rate: 0.4761905

Machine Learning Models:

1) Support Vector Machine (SVM):

```
Parameters:
                 SVM-Type:  C-classification
                 SVM-Kernel:  radial
                cost:  1
                gamma:  0.1428571
                Number of Support Vectors:  946
                Number of Classes:  2
                Levels:  Down Up
```

Confusion Matrix

	Actual	
Model	Down	Up
Down	1	0
Up	110	141

Misclassification rate: 0.4365079

2) Random Forest:

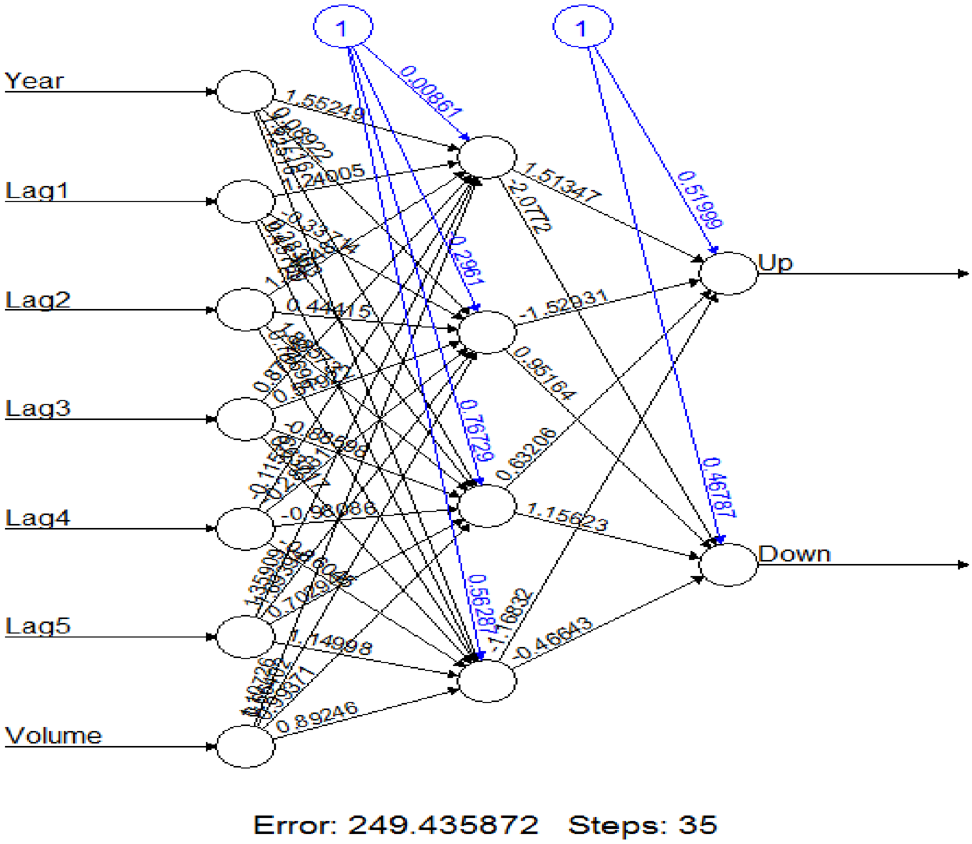
```
summary(rfmodel)
Type of random forest: classification
Number of trees: 500
No. of variables tried at each split: 2
 OOB estimate of  error rate: 49.4%
rfmodel$importance
       MeanDecreaseGini
Year          22.85293
Lag1          83.01989
Lag2          77.37586
Lag3          77.67590
Lag4          77.00136
Lag5          78.16711
Volume        82.25598
```

Confusion Matrix

	Actual	
Model	Down	Up
Down	64	38
Up	67	83

Misclassification rate(Test): 0.4166

3) Neural Net Model:



Confusion Matrix

	Actual	
Model	Down	Up
Down	109	137
Up	2	4

Misclassification rate(Test): 0.5515

Conclusion:
The results are intriguing. The best performing Models among Statistical Models is the Logistic Model with lowest misclassification error rate.

Among machine Learning Models Learning Models, Random Forest is the best predictor with lowest Misclassification error.
Machine Learning Algorithm Prediction completely depends on the data, sample size and latent construct.

The Research Team,

ALBEDO ENERGY
##