Stock market data analysis -

Shubham Gupta
Analytics club head, CFI, IIT Madras

Objective: To predict if an unnamed stock will go up or down in one hour.

Dataset description: It has 609 explanatory variables which include stock prices, sectoral data, economic data, experts predictions and indexes. All observations are taken at 5 minutes interval. The binary variable to be predicted is called *'TargetVariable'* which indicates whether the unnamed stock will go up or down in one hour.

Dataset used: We have used the dataset from kaggle competitions website, current dataset is available at this link: https://www.kaggle.com/c/informs2010/data

Techniques/Models used:

- Logistic Regression models:
- with all independent variables
- Stepwise variable selection and most correlated variables
- L2 regularization

SVM models:

- Using Linear kernel
- Using Radial kernel

Results and evaluation criteria:

- Different Models are compared based on:
- Area under curve of ROC plots
- Sensitivity, Specificity and Accuracy values based on True positive rate and True negative rate taken from Confusion Matrix
- Balanced Classification Rate calculated from Confusion Matrix
- Average accuracy on performing k-fold cross validation

Results obtained:

- Logistic Regression: accuracy- 0.89 , AUC of ROC plot- 0.94
- SVM: accuracy- 0.92 , AUC of ROC plot- 0.96