

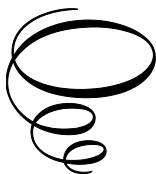
Machine Learning in the Analysis and Forecasting of Financial Time Series

Machine Learning in the Analysis and Forecasting of Financial Time Series

Edited by

Jaydip Sen and Sidra Mehtab

**Cambridge
Scholars
Publishing**



Machine Learning in the Analysis and Forecasting of Financial
Time Series

Edited by Jaydip Sen and Sidra Mehtab

This book first published 2022

Cambridge Scholars Publishing

Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

Copyright © 2022 by Jaydip Sen, Sidra Mehtab and contributors

All rights for this book reserved. No part of this book may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior permission of the copyright owner.

ISBN (10): 1-5275-8324-4

ISBN (13): 978-1-5275-8324-5

Dedicated to my sister Nabanita who left us on July 27, 2021.

—Jaydip

TABLE OF CONTENTS

List of Figures.....	ix
List of Tables.....	xiii
Preface.....	xvii
Chapter 1	1
Stock Price Prediction using Deep Learning and Natural Language Processing	
Jaydip Sen & Sidra Mehtab	
Chapter 2	29
Machine Learning and Deep Learning in Stock Price Prediction	
Jaydip Sen	
Chapter 3	68
Stock Price Prediction using Convolutional Neural Networks	
Jaydip Sen & Sidra Mehtab	
Chapter 4	102
Robust Predictive Models for the Indian IT Sector using Machine Learning and Deep Learning	
Abhishek Dutta & Jaydip Sen	
Chapter 5	197
A Causality Analysis between the Indian Information Technology Sector Index and the DJIA Index	
Ashmita Paul & Jaydip Sen	
Chapter 6	235
Stock Price Prediction using Machine Learning and Deep Learning Algorithms and Models	
Sidra Mehtab & Jaydip Sen	

Chapter 7	304
Analysis of Different Sectors of the Indian Economy for Robust Portfolio Construction Jaydip Sen	
Contributors.....	361

LIST OF FIGURES

Figure 2-1. The ANN model for classification (<i>Case II</i>).....	47
Figure 2-2. The decision tree regression model (<i>Case III</i>)	52
Figure 2-3. The loss convergence in the LSTM model (<i>Case III</i>).....	56
Figure 3-1. The architecture of the CNN#1 model.....	79
Figure 3-2. The architecture of the CNN#2 model.....	81
Figure 3-3. The architecture of the CNN#3 model.....	83
Figure 3-4. The architecture of the CNN#4 model.....	84
Figure 3-5. The variation of RMSE for CNN#1 model.....	90
Figure 3-6. The variation of RMSE for CNN#2 model.....	91
Figure 3-7. The variation of RMSE for CNN#3 model.....	92
Figure 3-8. The variation of RMSE for CNN#4 model.....	93
Figure 4-1(a). Bagging – actual and predicted values (<i>Case I</i>)	114
Figure 4-1(b). Bagging- actual and predicted relationship (<i>Case I</i>)	114
Figure 4-2. Bagging – actual and predicted values (<i>Case II</i>).....	115
Figure 4-3 (a). Bagging – the residual plot (<i>Case I</i>).....	116
Figure 4-3 (b). Bagging – the residual plot (<i>Case II</i>)	116
Figure 4-4 (a). Boosting – actual and predicted values (<i>Case I</i>).....	118
Figure 4-4 (b). Boosting – actual and predicted relationship (<i>Case I</i>).....	119
Figure 4-5. Boosting – actual and predicted relationship (<i>Case II</i>)	119
Figure 4-6 (a). Boosting – the residual plot (<i>Case I</i>).....	120
Figure 4-6 (b). Boosting – the residual plot (<i>Case II</i>).....	120
Figure 4-7 (a). The decision tree regression model (<i>Case I</i>)	122
Figure 4-7 (b). Decision tree – actual and predicted values (<i>Case I</i>).....	123
Figure 4-7 (c). Decision tree – actual and predicted relation (<i>Case I</i>).....	123
Figure 4-8 (a). Decision tree – actual and predicted values (<i>Case II</i>).....	125
Figure 4-8 (b). Decision tree – actual and predicted relation (<i>Case II</i>) ..	125
Figure 4-9 (a). Decision tree – the residual plot (<i>Case I</i>)	126
Figure 4-9 (b). Decision tree – the residual plot (<i>Case II</i>).....	126
Figure 4-10 (a). Random forest – actual and predicted values (<i>Case I</i>) ..	128
Figure 4-10 (b). Random forest – act and pred relation (<i>Case I</i>).....	128
Figure 4-11. Random forest – actual and predicted (<i>Case II</i>).....	130
Figure 4-12(a). Random forest – the residual plot (<i>Case I</i>).....	131
Figure 4-12(b). Random forest – the residual plot (<i>Case II</i>)	131
Figure 4-13. The architecture of the ANN regression model (<i>Case I</i>).....	133
Figure 4-14 (a). ANN – actual and predicted values (<i>Case I</i>)	134

Figure 4-14 (b). ANN – actual and predicted relation (<i>Case I</i>)	134
Figure 4-15. The architecture of the ANN regression model (<i>Case II</i>) ...	135
Figure 4-16. ANN regression– actual and predicted values (<i>Case II</i>)	136
Figure 4-17 (a). ANN regression – the residual plot (<i>Case I</i>).....	137
Figure 4-17 (b). ANN regression – the residual plot (<i>Case II</i>).....	137
Figure 4-18 (a). Multivariate regression – act and pred values (<i>Case I</i>)..	140
Figure 4-18 (b). Multivariate reg – act. and pred. relation (<i>Case I</i>).....	140
Figure 4-19. Multivariate regression – act. and pred. values (<i>Case II</i>)....	141
Figure 4-20 (a). Multivariate regression – the residual plot (<i>Case I</i>).....	142
Figure 4-20 (b). Multivariate regression – the residual plot (<i>Case II</i>).....	142
Figure 4-21. MARS regression – post model tuning (<i>Case I</i>)	145
Figure 4-22 (a). MARS – actual and predicted values (<i>Case I</i>).....	147
Figure 4-22 (b). MARS – actual and predicted relation (<i>Case I</i>)	146
Figure 4-23. MARS – actual and predicted values (<i>Case II</i>).....	147
Figure 4-24 (a). MARS regression – the residual plot (<i>Case I</i>).....	148
Figure 4-24 (b). MARS regression – the residual plot (<i>Case II</i>)	148
Figure 4-25. SVM model - a hyperplane separating two classes.....	150
Figure 4-26 (a). SVM regression – act and pred values (<i>Case I</i>).....	151
Figure 4-26 (b). SVM – actual and predicted relation (<i>Case I</i>)	151
Figure 4-27. SVM regression – act and pred values (<i>Case II</i>).....	153
Figure 4-28 (a). SVM regression – the residual plot (<i>Case I</i>).....	154
Figure 4-28 (b). SVM regression – the residual plot (<i>Case II</i>).....	154
Figure 4-29 (a). Logistic regression – lift curve (<i>Case I</i>)	159
Figure 4-29 (b). Logistic regression – ROC curve (<i>Case I</i>).....	160
Figure 4-30 (a). Logistic regression – act and pred prob (<i>Case II</i>).....	162
Figure 4-30 (b). Logistic regression – lift curve (<i>Case II</i>).....	163
Figure 4-30 (c). Logistic regression – ROC curve (<i>Case II</i>)	164
Figure 4-31 (a). Bagging – actual and predicted probabilities (<i>Case I</i>)...	166
Figure 4-31 (b). Bagging – actual and predicted probabilities (<i>Case II</i>)..	167
Figure 4-32. KNN – actual and predicted probabilities (<i>Case I</i>).....	169
Figure 4-33. Decision tree classification model (<i>Case I</i>).....	171
Figure 4-34. Architecture of the ANN classification model (<i>Case I</i>).....	173
Figure 4-35 (a). ANN – actual and predicted probs (<i>Case I</i>).....	174
Figure 4-35 (b). ANN – actual and predicted probs (<i>Case II</i>)	174
Figure 4-36. Boosting – actual and predicted probs (<i>Case I</i>)	177
Figure 4-37 (a). Random forest – act and pred probabilities (<i>Case I</i>)	179
Figure 4-37 (b). Random forest – actual and predicted probs (<i>Case II</i>) ..	179
Figure 4-38 (a). SVM – actual and predicted probs (<i>Case I</i>).....	181
Figure 4-38 (b). SVM – actual and predicted probabilities (<i>Case II</i>).....	181
Figure 4-39. Indian IT sector daily index values for the training LSTM ...	184
Figure 4-40. The convergence of the loss for the LSTM model.....	185

Figure 4-41. Training and validation accuracies of the LSTM model.....	186
Figure 4-42. The architecture of the LSTM regression model	187
Figure 4-43. Actual vs the predicted values of the LSTM model.....	188
Figure 5-1. Multivariate IT_DJIA – act and pred values (<i>Case I</i>)	207
Figure 5-2. Multivariate IT_DJIA – act and pred values (<i>Case II</i>)	207
Figure 5-3. MARS regression IT_DJIA – act vs. pred values (<i>Case I</i>)....	209
Figure 5-4. MARS regression IT_DJIA – act vs. pred values (<i>Case II</i>) ..	210
Figure 5-5. Bagging IT_DJIA – act and pred values (<i>Case I</i>).....	211
Figure 5-6. Bagging IT_DJIA – act and pred values (<i>Case II</i>)	212
Figure 5-7. Boosting IT_DJIA – act and the pred values (<i>Case I</i>).....	213
Figure 5-8. Boosting IT_DJIA – act and pred relation (<i>Case I</i>).....	213
Figure 5-9. Boosting IT_DJIA – the residual plot (<i>Case I</i>).....	213
Figure 5-10. Boosting IT_DJIA – act and pred values (<i>Case II</i>)	214
Figure 5-11. Boosting IT_DJIA – act and pred relation (<i>Case II</i>).....	214
Figure 5-12. Boosting IT_DJIA –the residual plot (<i>Case II</i>)	214
Figure 5-13. Random forest IT_DJIA – act and pred values of (<i>Case I</i>)..	216
Figure 5-14. Random forest IT_DJIA – act and pred relation (<i>Case I</i>)....	216
Figure 5-15. Random forest IT_DJIA –the residual plot (<i>Case I</i>)	216
Figure 5-16. Random forest IT_DJIA – act and pred values (<i>Case II</i>)	217
Figure 5-17. Random forest IT_DJIA – act and pred relation (<i>Case II</i>) ...	217
Figure 5-18. Random forest IT_DJIA – the residual plot (<i>Case II</i>)	217
Figure 5-19. Architecture of ANN regression model IT_DJIA (<i>Case I</i>)..	219
Figure 5-20. ANN IT_DJIA – actual and the predicted values (<i>Case I</i>)..	219
Figure 5-21. ANN IT_DJIA – act and the pred values (<i>Case II</i>)	220
Figure 5-22. ANN regression model architecture IT	220
Figure 5-23. SVM IT_DJIA model – act and the pred values (<i>Case I</i>)....	222
Figure 5-24. SVM IT_DJIA model – act and the pred values (<i>Case II</i>) ..	222
Figure 5-25. Indian IT sector index for training IT_DJIA (<i>Case I</i>)	225
Figure 5-26. The IT_DJIA LSTM model loss (<i>Case I</i>).....	226
Figure 5-27. The IT_DJIA LSTM model loss (<i>Case II</i>)	226
Figure 5-28. The IT LSTM model loss (<i>Case I</i>).....	227
Figure 5-29. The IT LSTM model loss (<i>Case II</i>).....	227
Figure 6-1. The logistic regression – act vs. pred probs (<i>Case III</i>).....	259
Figure 6-2. The logistic regression–the <i>lift</i> curve (<i>Case III</i>).....	259
Figure 6-3. The logistic regression – the <i>ROC curve</i>	259
Figure 6-4. The decision tree for classification (<i>Case III</i>).....	262
Figure 6-5. The bagging classification – act vs. pred probs (<i>Case III</i>)....	263
Figure 6-6 The boosting classification – act vs pred probs (<i>Case III</i>)	265
Figure 6-7. The ANN classification model (<i>Case III</i>)	268
Figure 6-8. The ANN classifier – act vs pred probs (<i>Case III</i>).....	268
Figure 6-9. Multivariate regression – act and pred values (<i>Case III</i>)	271

Figure 6-10. Multivariate regression – act and pred relation (<i>Case III</i>) ..	271
Figure 6-11. Multivariate regression – the residual plot (<i>Case III</i>)	272
Figure 6-12. The MARS regression – act and pred values (<i>Case III</i>).....	274
Figure 6-13. The MARS regression – act and pred relation (<i>Case III</i>)....	275
Figure 6-14. The MARS regression – the residual plot (<i>Case III</i>)	275
Figure 6-15. Decision tree regression model (<i>Case III</i>)	276
Figure 6-16. Decision tree – act and pred values (<i>Case III</i>)	277
Figure 6-17. Decision tree – act and pred relation (<i>Case III</i>)	277
Figure 6-18. Decision tree – the residual plot (<i>Case III</i>)	277
Figure 6-19. Bagging – actual and predicted values (<i>Case III</i>)	279
Figure 6-20. Bagging – act and pred relation (<i>Case III</i>).....	279
Figure 6-21. Bagging – the residual plot (<i>Case III</i>).....	279
Figure 6-22. Boosting – actual and predicted values (<i>Case III</i>)	281
Figure 6-23. Boosting – act and pred relation (<i>Case III</i>).....	281
Figure 6-24. Boosting – the residual plot (<i>Case III</i>).....	282
Figure 6-25. Random forest – act and pred values (<i>Case III</i>).....	283
Figure 6-26. Random forest – act and pred relation (<i>Case III</i>).....	283
Figure 6-27. Random forest – the residual plot (<i>Case III</i>).....	284
Figure 6-28. ANN regression model architecture (<i>Case III</i>)	285
Figure 6-29. ANN regression – act and pred values (<i>Case III</i>)	285
Figure 6-30. ANN regression – act and pred relation (<i>Case III</i>)	286
Figure 6-31. ANN regression – the residual plot (<i>Case III</i>)	286
Figure 6-32. SVM regression – act and pred values (<i>Case III</i>)	287
Figure 6-33. SVM regression – act and pred relation (<i>Case III</i>)	288
Figure 6-34. SVM regression – the residual plot (<i>Case III</i>)	288
Figure 6-35. The stock data of Godrej consumers for 2013 (<i>Case I</i>)	289
Figure 6-36. LSTM model architecture (<i>Cases I, II, and III</i>)	289
Figure 6-37. LSTM regression model loss (<i>Case I</i>)	290
Figure 6-38. The stock data of Godrej Consumers for 2014 (<i>Case II</i>)	291
Figure 6-39. LSTM model loss (<i>Case II</i>)	291
Figure 6-40. Godrej Consumers stock price for 2013-14 (<i>Case III</i>).....	292
Figure 6-41. LSTM model loss (<i>Case III</i>)	292
Figure 7-1. Decomposition of Indian banking sector index time series ..	314
Figure 7-2. Actual and predicted Indian IT index using <i>Method I</i>	340
Figure 7-3. Actual and predicted Indian IT index using <i>Method II</i>	342
Figure 7-4. First-order difference of the Indian IT index series	345
Figure 7-5. The ACF plot of the first-order difference series.....	346
Figure 7-6. The PACF plot of the first-order difference series.....	346
Figure 7-7. The ACF plot of the residuals for the best-fit ARIMA	348
Figure 7-8. Actual and predicted Indian IT index using <i>Method III</i>	349
Figure 7-9. Actual and predicted IT sector index using Method IV	351

LIST OF TABLES

Table 1-1. Logistic regression classification results	16
Table 1-2. KNN classification results.....	16
Table 1-3. Decision tree (CART) classification results	16
Table 1-4. Bagging classification results.....	16
Table 1-5. Boosting classification results.....	17
Table 1-6. Random forest classification results.....	17
Table 1-7. ANN classification results.....	17
Table 1-8. SVM classification results.....	17
Table 1-9. Multivariate regression results	18
Table 1-10. Decision tree regression results.....	18
Table 1-11. Bagging regression results	18
Table 1-12. Boosting regression results.....	18
Table 1-13. Random forest regression results	18
Table 1-14. ANN regression results	19
Table 1-15. SVM regression results	19
Table 1-16. LSTM regression results	19
Table 1-17. Granger test results at different lags.....	20
Table 1-18. Performance of sentiment augmented model	20
Table 2-1. Logistic regression classification results.....	41
Table 2-2. KNN classification results.....	42
Table 2-3. Decision tree (CART) classification results	42
Table 2-4. Bagging classification results.....	44
Table 2-5. Boosting (AdaBoost) classification results	44
Table 2-6. Random forest classification results.....	46
Table 2-7. ANN classification results.....	46
Table 2-8. SVM classification results.....	48
Table 2-9. Multivariate regression results	50
Table 2-10. MARS regression results.....	50
Table 2-11. Decision tree regression results.....	51
Table 2-12. Bagging regression results	53
Table 2-13. Boosting regression results.....	53
Table 2-14. Random forest regression results	54
Table 2-15. ANN regression results	54
Table 2-16. SVM regression results	55
Table 2-17. LSTM regression results	57

Table 2-18. The best performing classification models.....	58
Table 2-19. The best performing regression models	58
Table 3-1. Logistic regression classification results.....	86
Table 3-2. KNN classification results.....	86
Table 3-3. Decision tree (CART) classification results	86
Table 3-4. Bagging Classification results	86
Table 3-5. Boosting classification results.....	87
Table 3-6. Random forest classification results.....	87
Table 3-7. ANN classification results.....	87
Table 3-8. SVM classification results.....	87
Table 3-9. Multivariate regression results	88
Table 3-10. Decision tree regression results.....	88
Table 3-11. Bagging regression results	88
Table 3-12. Boosting regression results.....	88
Table 3-13. Random forest regression results	88
Table 3-14. ANN regression results	88
Table 3-15. SVM regression results	89
Table 3-16. The performance results of CNN#1 model	89
Table 3-17. The performance results of CNN#2 model	90
Table 3-18. The performance results of CNN#3 model	91
Table 3-19. The performance results of CNN#4 model	93
Table 3-20. The best performing ML classification models.....	93
Table 3-21. The best performing ML regression models	93
Table 4-1. Bagging regression results	117
Table 4-2. Boosting regression results.....	121
Table 4-3. Decision tree regression results.....	127
Table 4-4. Random forest regression results	132
Table 4-5. ANN regression results	138
Table 4-6. Multivariate regression results	143
Table 4.7. MARS regression results.....	149
Table 4.8. SVM regression results.....	155
Table 4.9. Logistic regression classification results	165
Table 4-10. Bagging classification results.....	168
Table 4-11. KNN classification results.....	170
Table 4-12. Decision tree classification results	172
Table 4-13. ANN classification results.....	175
Table 4-14. Boosting classification results.....	178
Table 4-15. Random forest classification results.....	180
Table 4-16. SVM classification results.....	182
Table 4-17. The performance results of the LSTM regression model	189
Table 5-1. Granger causality test with DJIA index as a predictor	205

Table 5-2. Multivariate regression results	208
Table 5-3. MARS regression results.....	210
Table 5-4. Bagging regression results	212
Table 5-5. Boosting regression results.....	215
Table 5-6. Random forest regression results	218
Table 5-7. ANN regression results	221
Table 5-8. SVM regression results	223
Table 5-9. Summary of the IT_DJIA regression models (<i>Case II</i>).....	228
Table 5-10. Summary of the IT regression models (<i>Case II</i>) of IT	228
Table 6-1. The performance results of the logistic regression models	258
Table 6-2. The performance results of the KNN models.....	260
Table 6-3. The performance results of the decision tree classifiers.....	261
Table 6-4. The performance results of the bagging classifiers	263
Table 6-5. The performance results of the boosting classifiers	264
Table 6-6. The performance results of the random forest classifiers.....	266
Table 6-7. The performance results of the ANN classifiers	267
Table 6-8. The performance results of the SVM classifiers	268
Table 6-9. The performance of the multivariate regression models	271
Table 6-10. The performance results of the MARS regression models ...	274
Table 6-11. The performance of the decision tree regression models	278
Table 6-12. The performance of the bagging regression models.....	280
Table 6-13. The performance of the boosting regression models.....	281
Table 6-14. The performance of the random forest regression.....	283
Table 6-15. The performance of the ANN regression models.....	284
Table 6-16. The performance of the SVM regression models.....	286
Table 6-17. Summary of the classification models (<i>Case I</i>).....	293
Table 6-18. Summary of the classification models (<i>Case II</i>)	293
Table 6-19. Summary of the classification models (<i>Case III</i>)	294
Table 6-20. Summary of the regression models (<i>Case I</i>)	295
Table 6-21. Summary of the regression models (<i>Case II</i>).....	295
Table 6-22. Summary of the regression models (<i>Case III</i>).....	295
Table 7-1. Indian banking sector time series (Jan 2010 – Dec 2020).....	317
Table 7-2. Seasonality of different sectors (Jan 2010 – Dec 2020).....	321
Table 7-3. Price movements of the auto sector stocks.....	322
Table 7-4. Price movements of the banking sector stocks.....	323
Table 7-5. Price movements of the capital goods sector stocks.....	324
Table 7-6. Price movements of the consumer durable sector stocks	325
Table 7-7. Price movements of the FMCG sector stocks	326
Table 7-8. Price movements of the healthcare sector stocks	327
Table 7-9. Price movements of the IT sector stocks.....	328
Table 7-10. Price movements of the large cap sector stocks	329

Table 7-11. Price movements of the metal sector stocks.....	330
Table 7-12. Price movements of the mid cap sector stocks.....	331
Table 7-13. Price movements of the oil and gas sector stocks	332
Table 7-14. Price movements of the power sector stocks.....	333
Table 7-15. Price movements of the realty sector stocks.....	334
Table 7-16. Price movements of the small cap sector stocks.....	335
Table 7-17. Price movements of the telecom sector stocks	336
Table 7-18. Percentage of cases matching sectoral seasonality.....	336
Table 7-19. Forecasted Indian IT sector index using <i>Method I</i>	340
Table 7-20. Forecasted Indian IT sector index using <i>Method II</i>	341
Table 7-21. The output of the <i>auto.arima</i> (Jan 2010 – Dec 2019).....	343
Table 7-22. The six candidate ARIMA models	347
Table 7-23. Forecasted Indian IT sector index using <i>Method III</i>	348
Table 7-24. Iterations of building the model for <i>Method IV</i>	350
Table 7-25. Forecasted Indian IT sector index using <i>Method IV</i>	351
Table 7-26. Performance of the four methods of forecasting	352

PREFACE

The financial sector including financial services, banking, and insurance has witnessed the maximum applications and use cases of machine learning, deep learning, and artificial intelligence. While the financial organizations have only skimmed the surface of the rapidly evolving areas such as deep neural networks and reinforcement learning, the possibility of applying such techniques in many applications in finance and econometrics vastly remains unexplored yet.

The chapters in the volume present several techniques of financial time series analysis and forecasting financial series using statistical, econometric, machine learning, and deep learning approaches. The historical data of the daily and monthly index values of various sectors and important stocks listed in the National Stock Exchange (NSE) of India and the Bombay Stock Exchange (BSE) are used in building predictive models which are later on used to predict the future values of the index and their movement patterns. The time series decomposition results of the financial sectors provide the readers with several useful insights into the behavioral characteristics of different sectors which can prove important for understanding the sectors in a better way. A deeper understanding of the sectoral behavioral patterns will enable the investors to take more effective investment decisions and gain higher profits. The statistical and econometric modeling approaches discussed in the chapters include exponential smoothing methods, Holt and Winter trend and seasonality method, autoregressive (AR) and moving average (MA) method, autoregressive integrated moving average (ARIMA) method, Granger causality analysis, univariate linear regression, multivariate linear regression, and multivariate adaptive regression spline (MARS). The machine learning-based predictive models include several classification and regression approaches including logistic regression, k -nearest neighbors, decision trees, bagging, adaptive boosting, extreme gradient boosting, random forest, support vector machine, and artificial neural network. In the deep learning category, the applications of convolutional neural network (CNN) and long-and-short-term memory (LSTM) network architectures are demonstrated in designing predictive models for financial time series data. The use of text mining and natural language processing (NLP) in building precise financial models is also presented in one of the chapters in the volume.

In Chapter 1, *Stock Price Prediction using Deep Learning and Natural Language Processing*, Sen & Mehtab propose several machine learning and deep learning-based models for predicting NIFTY 50 index values and their movement patterns on the National Stock Exchange (NSE) of India. The models are trained on historical NIFTY index values from January 4, 2010, to December 31, 2018. The evaluation of the models is done on test data from January 1, 2019, to December 31, 2019. The authors present an additional text analysis module to improve the accuracy of the predictive models. This is done by incorporating a sentiment analysis module that analyses public sentiment on Twitter on the NIFTY 50 stocks. The output of the sentiment analysis module is used as the second input to the predictive model in addition to the past NIFTY 50 index values. To model the nonlinear relationship between Twitter sentiment and the NIFTY index values at different lags, the authors design a five-layer self-organizing fuzzy neural network (SOFNN) that uses the ellipsoidal function as the basis function. The results indicate that the use of the sentiment analysis module further increases the prediction accuracy of the deep learning-based predictive model.

In Chapter 2, *Machine Learning and Deep Learning in Stock Price Prediction*, Sen presents several machine learning-based models for predicting stock price movement in a short-term time frame using classification techniques. The stock prices are also predicted using a short forecast horizon using several regression models as well. A deep learning-based regression built on long-and-short-term memory (LSTM) network architecture is also presented. The classification and regression models are trained and tested on the historical prices of two important stocks listed on the NSE of India. Extensive results are provided on the performances of the models. The results clearly show that the LSTM is far superior to its machine learning counterparts in predicting future stock prices and their movement patterns for stock price records with rich features and high frequencies.

In Chapter 3, *Stock Price Prediction using Convolutional Neural Networks*, Sen & Mehtab use daily index values of NIFTY 50 from January 29, 2014, to July 31, 2020, for training and testing several machine learning-based classification and regression models. The models are trained on data from December 29, 2014, to December 28, 2018, while evaluation of the models was done on data from December 31, 2018 to July 31, 2020. The classification models are used for predicting the movement patterns of the daily NIFTY index during the test period. On the other hand, the daily NIFTY index values are predicted using the regression models. Four deep learning-based regression models built on the convolutional neural network (CNN) architecture are also proposed. The authors have presented detailed

results on the performances of the predictive models. The results have indicated that the CNN models are superior to the machine learning-based regression models in terms of their prediction accuracies.

In Chapter 4, *Robust Predictive Models for the Indian IT Sector using Machine Learning and Deep Learning*, Dutta & Sen present several machine learning models of classification and regression for accurately predicting the future daily index values and their movement patterns of the Indian information technology (IT) sector listed on the NSE. The Indian IT sector daily index values on the NSE from January 2008 to December 2018 are used for training the models. The models are tested on the data from January 2019 to December 2019. The authors present an extensive set of results that indicates the deep learning model built on LSTM architecture is more accurate than its machine learning counterparts in its future prediction of the daily IT index.

In Chapter 5, *A Causality Analysis between the Indian Information Technology Sector Index and the DJIA Index*, Paul & Sen investigate causality between the information technology sector index on the NSE of India and the Dow Jones Industrial Average (DJIA) index of the USA. In their work, the authors use the historical index values from January 5, 2009, to December 27, 2019. The historical index values from January 5, 2009, to December 28, 2018, are used for training the models while testing is done on the data from December 31, 2018, to December 27, 2019. Two regression models are built for predicting the index values of the Indian IT sector on the NSE. The first one is an autoregressive model that uses the lag values of the series as the predictors, while the other one is multivariate. The multivariate model uses the lag values of the DJIA index in addition to the lag values of the Indian IT sector index as the predictors. For both models, the target variable is the Indian IT sector index. It is found that the multivariate model is more accurate indicating a causal effect of the DJIA index series on the Indian IT sector index series.

In Chapter 6, *Stock Price Prediction using Machine Learning and Deep Learning Algorithms and Models*, Mehtab & Sen demonstrate how stock prices and stock returns and their movement patterns can be predicted on a short-term horizon with a fairly high level of accuracy. The authors propose a composite framework consisting of several statistical, econometric, machine learning, and deep learning architectures and models for predicting the future prices and the future price movements of an important stock listed on the NSE of India. The stock price data consist of records collected at five minutes intervals. The records are aggregated into three slots in a day, and the models are used to predict the price and its movement for successive slots. A deep learning regression model is also

designed on the LSTM architecture and its prediction accuracies are compared with those of the machine learning models. The results show that while all the models are reasonably accurate in their prediction, the LSTM model outperforms the machine learning model yielding a very high level of precision in forecasting.

In Chapter 7, *Analysis of Different Sectors of the Indian Economy for Robust Portfolio Construction*, Sen analyzes the behavioral patterns of the daily index values of fifteen sectors of the Indian stock market listed on the Bombay Stock Exchange (BSE). The author uses the time series decomposition approach to decompose the daily index series of the sectors from January 2010 to December 2020 into their trend, seasonality, and random components. Based on the decomposition results, the author illustrates how the sectors exhibit a difference in their behavior in trend, seasonality, and randomness, and the way these differences can be utilized by the investors to gain profit from the stock market. Furthermore, four methods of forecasting are proposed for predicting monthly sectoral index, and their prediction accuracies are compared. These forecasting models will help the analysts and investors in predicting the future index values accurately.

While the chapters in this volume do not primarily deal with the basic theories on the topics involved, all the relevant principles and fundamentals are discussed in brief in the chapters for the sake of completeness. Hence, even if some background knowledge of statistics, econometrics, and machine learning may be useful, the book does not presume it for the readers. We believe that the volume can be a valuable resource to anybody interested in gaining knowledge in financial time series analysis. However, the primary target audience for the book is the advanced postgraduate and doctoral students of finance, management, data science, computer science, information technology, and econometrics. In addition, faculty members of graduate schools and universities, practitioners in the industry working in the area of financial analytics, risk management, security analysis, and portfolio management, will also find the subject matters discussed in the book quite useful.

We express our sincere thanks to all the contributors to the chapters in the volume. Without their valuable contributions, it would have been impossible to make this project a success. We also express our sincere thanks to Cambridge Scholars Publishing for providing us with the opportunity to publish our work with their prestigious publishing house. Special thanks are due to Adam Rummens and Amanda Miller of Cambridge Scholars Publishing for their patience, cooperation, and support during different phases of the long publishing process. The members of our

respective families and our colleagues have been always the sources of our inspiration, and motivation during such a long, painstaking, and complex scholastic project. Without their support, the publication of this volume would not have been possible. Many thanks to all of them!

Jaydip Sen & Sidra Mehtab
Editors

CHAPTER 1

STOCK PRICE PREDICTION USING DEEP LEARNING AND NATURAL LANGUAGE PROCESSING

JAYDIP SEN & SIDRA MEHTAB

Introduction

Prediction of the future movement of stock prices has been the subject of many studies. On the one hand, there are proponents of the *efficient market hypothesis* (EMH) who claim that stock prices cannot be predicted. Alternatively, some studies have shown that, if correctly modeled, stock prices can be predicted with a reasonable degree of accuracy. The latter focuses on the choice of variables, appropriate functional forms, and forecasting techniques. In this regard, Sen and Datta Chaudhuri proposed a novel approach to stock price forecasting based on a time series decomposition of the stock price time series (Sen & Datta Chaudhuri, 2016a; Sen & Datta Chaudhuri, 2016b; Sen & Datta Chaudhuri, 2016c; Sen & Datta Chaudhuri, 2016d; Sen & Datta Chaudhuri, 2017a; Sen & Datta Chaudhuri, 2017b; Sen & Datta Chaudhuri, 2018; Sen, 2017a; Sen, 2017b). Propositions also exist that use a granular approach to stock price prediction in a short-term time frame using machine learning- and deep learning-based models (Sen & Datta Chaudhuri, 2017c; Sen, 2018).

There are propositions in the literature on the technical analysis of stock prices where the objective is to identify patterns in stock movements and derive profits from them. Various indicators such as *Bollinger Bands*, *moving average convergence divergence* (MACD), *relative strength index* (RSI), *moving average* (MA), *momentum stochastics* (MS), and *meta sine wave* (MSW) have been devised towards this end. For gaining profits, traders extensively use patterns such as *head and shoulders*, *triangle*, *flag*, *Fibonacci fan*, and *Andrew's pitchfork*. These approaches provide the user

with visual manifestations of the indicators, which help ordinary investors understand how stock prices may move.

This chapter proposes several machine learning and deep learning-based predictive models for predicting the NIFTY 50 index movement in the National Stock Exchange (NSE) of India. We use the daily stock price values for the period January 4, 2010, to December 31, 2018, as the training dataset for building the models, and apply the models to predict the daily stock price movement and actual closing value of the stock for the period January 1, 2019, to December 31, 2019. We further augment the predictive model by incorporating a sentiment analysis module that analyses public sentiment on Twitter on the NIFTY 50 stocks. The sentiment analysis module's output is used as the second input to the model with the historical NIFTY 50 data to predict future stock price movements. Following the approach proposed by Mittal and Goel, we classify the public sentiment on Twitter into four classes and study the causal effect of these sentiment classes on the NIFTY 50 stock price movement using the Granger Causality Test (Mittal & Goel, 2012).

We organize the chapter as follows. In the section titled *Problem Statement*, we explicitly define the problem at hand. The section titled *Related Work* provides a brief review of related work on stock price movement prediction. In the section titled *Methodology*, we describe our research methodology. Extensive results on the performance of the predictive models are presented in the section titled *Performance Results*. This section also describes all the predictive models built in this study and the results they have produced. Finally, in the section titled *Conclusion*, we conclude the paper.

Problem Statement

Our proposed method is based on collecting the NIFTY 50 index's historical values from India's NSE and developing a robust forecasting framework for future stock index movement. We contend that index values' daily movement patterns can be learned using powerful machine learning and deep learning-based approaches. The knowledge gained can be applied to predict the future price movements of stocks. In this study, we choose the prediction horizon as one week. We hypothesize that the learning-based approaches can be further augmented by sentiment analysis of social media data on Twitter so that stock price movements can be predicted with even higher accuracy. Here, we do not address the forecasting of the short-term movement of the stock price for intra-day traders. Instead, our analysis is more relevant to long-term investors.

At any point in time in the Indian economy, given the appetite of financial market players, including individuals, domestic institutions, and foreign financial institutions, there is a finite amount of funds deployed in the stock market. This fund is distributed among various stocks. Thus, if some stock prices rise, other stock prices should fall. Using our proposed approach, an investor will be able to predict the movement pattern of the NIFTY 50 index, which depicts the stock market sentiment in India.

Related Work

We can classify the existing propositions in the literature on stock price movements and stock price predictions into four broad categories based on the choice of variables, approaches, and techniques adopted in modeling. The first category includes approaches that use *simple regression techniques* on cross-sectional data (Asghar et al., 2019; Dutta et al., 2012; Roy et al., 2015; Zhong & Enke, 2017; Jaffe et al., 1989). These models do not yield accurate results because stock price movement is a highly nonlinear process. The propositions in the second category exploit time series models and techniques using *statistical and econometric methods* such as *exponential smoothing*, *autoregressive integrated moving average* (ARIMA), *Granger causality test*, *autoregressive distributed lag* (ARDL), and *generalized autoregressive conditional heteroscedasticity* (GARCH) to forecast stock prices (Du, 2012; Khandelwal et al., 2015; Ning et al., 2019; Pai & Lin, 2005; Wang et al., 2020). The third strand includes propositions using *machine learning*, *deep learning*, and *natural language processing* to predict stock returns (Mehtab & Sen, 2020a; Mehtab et al., 2020d; Mehtab & Sen, 2022; Sen & Mehtab, 2021a; Sen & Mehtab, 2021b; Mehtab & Sen, 2022; Sen et al., 2020; Yang et al., 2017; Wu et al., 2012; Ballings et al., 2015; Lv et al., 2019). The propositions of the fourth category are based on *hybrid models* built on machine learning and deep learning with inputs of historical stock prices and sentiments in news articles on the social web (Attigeri et al., 2015; Mittal & Goel, 2012; Medhat et al., 2014; Nam & Seong, 2019).

First, we briefly discuss some of the regression-based propositions existing in the literature.

Zhong and Enke propose methods for the future stock price and price movement prediction using *auto-regressive moving average* (ARMA), *auto-regressive integrated moving average* (ARIMA), *generalized autoregressive conditional heteroscedastic* (GARCH), and *smooth transition autoregressive* (STAR) models (Zhong & Enke, 2017). The authors further present another set of statistical methods involving multiple

input variables. The models of this category are *linear discriminant analysis* (LDA), *quadratic discriminant analysis* (QDA), and *multivariate regression analysis*. The authors forecast the daily direction of the S&P500 index return based on sixty predictor variables. Three methods of dimensionality reduction are first applied to the dataset. The methods used for dimensionality reduction are (i) *principal component analysis* (PCA), (ii) *fuzzy robust principal component analysis* (FRPCA), and (iii) *kernel-based principal component analysis* (KPCA). A neural network-based classification model is applied to the transformed data to forecast the direction of movement of future returns. It is observed that the PCA-ANN combination yields the highest level of classification accuracy among all the dimensionality reduction methods.

Dutta et al. propose a logistic regression model using financial ratios as the predictors to predict future stock returns (Dutta et al., 2012). This study attempts to classify the performance of companies as *good* or *bad* based on their one-year performance.

Roy et al. propose a penalty-based regression method – *least absolute shrinkage and selection operator* (LASSO) – to predict future stock prices (Roy et al., 2015). LASSO is an efficient linear regression model that avoids model overfitting and enables feature selection. The regression model is tested on the historical stock prices of the Goldman Sachs Group Inc. The performance of the model is found to be superior to that of another penalty-based regression method, the Ridge regression model.

In the following, we provide a brief description of some of the works in the literature based on econometric methods, such as ARIMA, Granger causality, and GARCH.

Du argues that because stock prices are complex nonlinear functions of many economic factors, it is necessary to design nonlinear models to achieve higher forecasting accuracy (Du, 2018). The author proposes a composite model integrating an ARIMA and a *backpropagation* (BP) *neural network* to forecast the Shanghai Securities Composition stock index. The integrated model's forecasting performance is then compared with those of the individual ARIMA and BP models. It is found that the prediction accuracy of the ARIMA-BP model is superior to that of the BP model, which in turn, is observed to be better than that of the linear ARIMA model.

Wang et al. propose a method of combining asymmetry and extreme volatility in stock prices to forecast volatility in future stock prices (Wang et al., 2020). The authors demonstrate that a shock has a highly significant effect on stock price volatility. It is also shown that volatility is affected by the asymmetry effect both in the long- and short-term. The model, when

tested on out-of-sample data, yields an improved accuracy compared to the GARCH-MIDAS model.

Khandelwal et al. present a scheme based on the *discrete wavelet transform* (DWT) for time series forecasting (Khandelwal et al., 2015). The proposed method separates the linear and nonlinear components in a given time series by applying the DWT. After the time series is decomposed, an ARIMA and an *artificial neural network* (ANN) model are separately used on the time series to reconstruct the original time series from its decomposed components. Once the reconstruction is achieved with a sufficiently high accuracy level, the model is tested on four real-world financial time series. The performance of the model on the out-of-sample data is found to be superior to those of the standalone ARIMA, ANN, and the hybrid ARIMA and ANN models proposed by Zhang (Zhang, 2003).

In the following, we discuss the salient features of some machine learning and deep learning-based propositions in the literature for stock price prediction.

Yang et al. present an ensemble of deep learning models for forecasting the Shanghai Composite Index and the SZSE Component Index of the Chinese stock market (Yang et al., 2017). Using the *backpropagation algorithm* to minimize the error and the *Adam* optimizer to optimize the *gradient descent algorithm*, a set of component networks is built. An ensemble of all these component networks is constructed using *bagging* to create a generalized model. The ensemble model is then tested on the out-of-sample data, and trend predictions are computed based on the predicted index values. The accuracy of the trend predictions of the daily *high* and *low* values for the Shanghai Composite Index is 74.15%. The corresponding values for the SZSE Component Index are 73.95% and 72.34%, respectively.

Wu et al. present a novel approach for predicting stock price trends using a combination of *sequential chart patterns*, *l-means*, and the *AprioriAll algorithm* (Wu et al., 2012). The time series of stock prices is divided into a set of subsequences. For each subsequence, appropriate charts are constructed using a *sliding window* method. The charts are then clustered using the *k-means* algorithm (Wu, 2012). After clustering, the charts form the chart pattern sequences, which are mined further to identify frequently occurring patterns using the *AprioriAll* algorithm (Yu & Li, 2018). The existence of such frequent patterns indicates that some market behaviors are exhibited in a correlated fashion. Detection of the correlations of such hidden market behaviors enables accurate prediction of the trend in stock price. The authors validate their propositions by experimental results.

Ballings et al. present a scheme that compares the performances of ensemble methods such as *random forest*, *AdaBoost*, and *kernel factory*,

with single classifier models such as *neural networks*, *logistic regression*, *support vector machines*, and *k-nearest neighbors* (Ballings et al., 2015). Using the time series of historical stock prices of 5767 public companies and the *area under the curve* (AUC) for the *receiver operating characteristic curve* (ROC) as the metric, the models' performances are compared over a forecast horizon of one year. The models, listed in the decreasing order of their performance, are *random forest*, *support vector machine*, *kernel factory*, *AdaBoost*, *neural network*, *k-nearest neighbors*, and *logistic regression*.

Mehtab and Sen present a series of works on the design of predictive models for predicting future stock prices and index values and movements using innovative machine learning and deep learning architectures (Mehtab & Sen, 2020a; Mehtab & Sen, 2020b; Mehtab & Sen, 2020c; Mehtab et al., 2020d; Mehtab & Sen, 2021a; Mehtab & Sen, 2021b; Mehtab & Sen, 2022). The authors use daily historical stock prices and stock index values at an interval of 5 minutes. Exploiting the power of *convolutional neural networks* (CNNs) and *long- and short-term memory* (LSTM) networks, the models are found to have achieved a high level of accuracy on the out-of-sample of data. The authors propose four CNN models and six LSTM models with different architectures and input data shapes (i.e., univariate or multivariate time series data). The models are compared on their *root mean square error* (RMSE) values. The results elicit two exciting observations: (i) the performances of the CNN models are superior to those of the LSTM models, and (ii) the models based on the univariate data yield more accurate results than the models using the multivariate data. In another set of works, Mehtab et al. propose further variants of CNN and LSTM-based models for predicting future stock price values and stock price movements (Mehtab et al., 2020d; Mehtab et al., 2021a; Sen & Mehtab, 2021b). The authors report extensive results of the performances of the models.

Finally, we discuss some of the hybrid models that use sentiments in the news and other relevant textual information for stock price prediction in the following.

Bollen et al. contend that emotions profoundly affect an individual's buy or sell decisions (Bollen et al., 2011). The authors propose a mechanism that computes the collective *mood* states of the public from many Twitter feeds and investigates whether the collective moods exhibit any correlation with the Dow Jones Industrial Average (DJIA) index. The public *moods* from the Twitter feed are computed using two mood tracking tools – (i) *OpinionFinder*, which classifies the *opinions* as either positive or negative, and (ii) the *Google Profile of Mood States* (GPOMS) which measures six dimensions of a *mood*. The six dimensions are – *calm*, *alert*, *sure*, *vital*,

kind, and *happy*. The results show that DJIA index values can be more accurately predicted by including specific public mood dimensions in the model. The model is further augmented using a *Granger causality* analysis and a *self-organizing fuzzy neural network* (SOFNN) so that the public moods can be more efficiently used as a predictor in forecasting the future DJIA index values. An accuracy level of 86% is found in predicting the daily up and down changes in the *DJIA index's close values* using the composite mode with the Granger causality and the SOFNN module.

Checkley et al. study the use of sentiment metrics extracted from microblogs in predicting stock market index values (Checkley et al., 2017). Using the bearish and bullish sentiments from micro-blogging sites at different time intervals, the authors model their forecasting ability in future stock price movements, price volatility, and trade volume. The study reveals a significant causal link between the sentiments on the micro-blogging sites to the price movements, volatility, and the volume traded at short-term time intervals. The study concludes that investors' behavior in stock markets is more like that of a hasty mob than a group of wise people.

Chen et al. study the impact of sentiments from the news sites on the stock market in Taiwan (Chen et al., 2019). After the news articles are collected from the web, they are preprocessed, and the text corpus is transformed into a word feature set using the *Word2Vec* approach (Jatnika et al., 2019). After the *Word2Vec* model is ready, an LSTM-based deep learning model predicts future stock prices using the news articles from the news sites.

Dang and Duong highlight the influence of online news articles on the movement patterns of stock prices (Dang & Duong, 2016). Based on their hypothesis, the authors propose using a time series analysis that integrates a gamut of text mining and text analysis mechanisms. The proposed model is tested on real-world time series of historical stock prices. The model is found to achieve an accuracy of around 73%.

Galvez and Gravano emphasize the impact of online sentiments on the news articles on the stock prices in Argentina (Galvez & Gravano, 2017). The authors investigate two critical issues. The first question that the authors pose is whether there is any useful information in the stock exchange boards for predicting stock returns. The second question is whether the information is novel or is carried in the time series of the historical stock prices. The authors train, validate, and test a series of predictive models using machine learning and topic modeling algorithms in text analysis to address these questions. The models are built using various combinations of features and predictor variables. The study reveals that the information extracted from the online exchange boards complements the information in the time series

of the historical stock prices. Hence, the use of information from the online exchange boards improves the performance of the predictive models.

Hu et al. argue that the quality, veracity, and comprehensiveness of online textual information related to the stock market and stock price movement vary widely, and in many cases, they are low-quality news (Hu et al., 2018). To make learning models robust against such chaotic information, the authors propose three principles of learning: (i) *sequential content dependency*, (ii) *diverse influence*, and (iii) *efficient and effective learning*. The proposition implements the first two characteristics of learning by including a *hybrid attention network* (HAN). The HAN attempts to predict the trend of the stock price time series using a sequence of recent and related online news. A controlled learning mechanism implements the third aspect of learning. Simulations show that trading strategies formulated based on the model yield a significantly improved annualized return on investment compared to the returns produced by the models without the text analytics module.

Jeong et al. propose a predictive model for forecasting stock prices and stock price movement that incorporates an *opinion mining* module (Jeong et al., 2018). The model is capable of performing three major tasks: (i) filtering fake information on the web, (ii) credit risk assessment, and (iii) detection of critical signals and using them in stock price prediction. These three operations are executed iteratively by the model on real-world time series of stock prices and news items available on the web. The results show that the proposed model helps an investor make investment decisions and increases the returns of investments.

Li et al. present a novel architecture of a predictive model for forecasting stock prices that uses an LSTM-based module and a text mining module that incorporates the sentiment of the investors and other market factors (Li et al., 2017). The model extracts the sentiment from the online news and opinions using a Naïve Bayes algorithm that filters out irrational and fake opinions. The proposed model is tested on the CSI300 index and produced an improved forecasting accuracy compared to the benchmark models that do not have any text processing components.

Li et al. discuss the limitation of the *bag-of-words* approach of online news processing in stock price prediction (Li et al., 2014). The authors argue that news sentiment should be considered a vital ring in the chain of sequence from the word patterns in the news to the final price movement of stocks. The authors use the Harvard psychological dictionary and the Loughran-McDonald financial sentiment dictionary to construct the vocabulary for the text processing module. After the vocabulary space is built, news articles are assigned quantitative measures and then placed in