Data Science in Theory and Practice

Data Science in Theory and Practice

Edited by

Jaydip Sen and Sayantani Roy Choudhury

Cambridge Scholars Publishing

Data Science in Theory and Practice

Edited by Jaydip Sen and Sayantani Roy Choudhury

This book first published 2024

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

ISBN: 978-1-0364-0898-5
ISBN (Ebook): 978-1-0364-0899-2

TABLE OF CONTENTS

PREFACE

The volume titled *Data Science in Theory and Practice* serves as a comprehensive exploration of the dynamic and rapidly evolving field of data science. In today's digital age, the proliferation of data from various sources has led to unprecedented opportunities and challenges. Organizations across industries are increasingly relying on data-driven insights to inform decision-making, drive innovation, and gain a competitive edge in the marketplace. At the heart of this data revolution lies the interdisciplinary field of data science, which combines principles from statistics, computer science, machine learning, and domain expertise to extract meaningful insights from vast and complex datasets. From financial markets to social dynamics, from agricultural landscapes to technological advancements, data science plays a pivotal role in unraveling insights and driving informed decision-making.

This edited volume aims to provide readers with a comprehensive understanding of both the theoretical foundations and practical applications of data science. Through a collection of chapters authored by experts in the field, the volume covers a wide range of topics, spanning from fundamental concepts to advanced methodologies and emerging trends. Each chapter offers a unique perspective on various aspects of data science, providing readers with valuable insights, practical guidance, and cutting-edge research findings. The book is organized into twelve chapters, each addressing various domains, techniques, and applications of data science.

In Chapter 1 titled *A Comparative Study of Portfolio Optimization Methods for the Indian Stock Market Sector Index*, Sen, Dasgupta, and Roy Choudhury present a comparative analysis of three portfolio optimization methods, mean-variance portfolio (MVP), hierarchical risk parity (HRP) portfolio, and hierarchical equal risk contribution (HERC) portfolio, in the Indian stock market across 15 sectors of stocks listed on the National Stock Exchange (NSE) of India. Top stocks in each sector are chosen based on their free-float market capitalization. Three portfolios per sector are constructed using data from July 1, 2019, to June 30, 2022, and evaluated from July 1, 2022, to June 30, 2023. Performance metrics include cumulative returns, volatility, and Sharpe ratios, identifying the portfolios with superior performance characteristics.

In Chapter 2 titled *A Portfolio Rebalancing Approach for the Indian Stock Market,* Sen, Dasgupta, Dasgupta, and Roy Choudhury introduce a calendar-based rebalancing approach to stock portfolios in the Indian market across ten sectors listed on the NSE of India. It selects the top ten stocks from each sector based on their market capitalization values. Using historical stock data from January 4, 2021, to September 20, 2023, portfolios are constructed and trained from January 4, 2021, to June 30, 2022. Performance evaluation is done from July 1, 2022, to September 20, 2023, with a focus on analyzing the effectiveness of the calendar rebalancing method across sectors.

In Chapter 3 titled *Technical Analysis of Indian Stocks: A Triad of Bollinger Bands, MACD, and RSI Strategies,* Sen, Roy Choudhury, and Pathak explore three widely-used technical indicators - Bollinger Bands, Moving Average Convergence Divergence (MACD), and Relative Strength Index (RSI) - and their efficacy in the Indian stock market. The proposed approach of the authors involves selecting 14 sectors listed on the National Stock Exchange (NSE) and identifying 10 top stocks from each sector based on their market capitalization. Trading activities are conducted from July 1, 2022, to June 30, 2023, with an initial capital of INR 100,000, using the three technical indicators. The performance of each indicator is evaluated, and a comparative analysis across all sectors is conducted to assess their effectiveness.

In Chapter 4 *Marital Status and Workforce Participation Decision of Indian Women*, Guha, Ghosh, Kapat, and Roy Choudhury examine the statistical significance of 'marital status' and several other associated factors influencing female workforce participation, with the Worker Population Ratio (WPR) serving as the primary variable of interest. Additionally, this study seeks to identify factors contributing to the proportion of NEET (Not in Employment, Education, and Training). The work shows that an analysis of these aspects provides insights into the factors affecting the employment status of women, encompassing both participation in the workforce and unemployment.

In Chapter 5 *Interstate Inequalities in LPG Consumption: A Study in Indian Context*, Koley and Majumdar present a study that focuses on focus on the state-wise LPG consumption status in India from 1955 until now. According to the authors, the state-wise LPG consumption trend in India reveals a massive inequality. While some states exhibit higher per capita consumption, there are states where the per capita LPG consumption has been significantly low. The work attempts to identify and focus on the factors responsible for interstate differences in LPG consumption in India over the last twenty years.

In Chapter 6 *MGNREGP of India and the Poor States: A Quest for Reality*, Tiwary and Khuntia examine the effectiveness of the social welfare scheme, Mahatma Gandhi National Rural Employment Guarantee Program (MGNREGP) in terms of person-days created and funding distributed to the states in India by the central government based on the needs of the states. The study also investigates how much benefits this scheme has been able to bring to the poorest in society.

In Chapter 7 *Ethical Practices in the Banking Sector: A Comparative Analysis of State Bank of India and HDFC Bank*, Dastidar and Osmani carry out a study on a comparative analysis of the perceptions of bank employees regarding ethical issues at two of the largest banks in India, the State Bank of India and the HDFC Bank. The authors investigate whether employees at these banks adhere to their respective codes of ethics in their job performance and analyze potential strategies for enhancing the ethical standards of the banks.

In Chapter 8 *A Comparative Study of Hyperparameter Tuning Methods*, Dasgupta and Sen present three well-known algorithms for hyperparameter tuning of models. Linear and non-linear models for regression and classification tasks based on some public datasets are built and their performances are evaluated. The non-linear models are found to outperform the linear models if their hyperparameters are tuned accurately.

In Chapter 9 *The Backpropagation Algorithm in Employee Competency Assessment*, Mukhopadhyay presents an implementation of the backpropagation algorithm for employee competency assessment. Competencies are categorized into various areas and defined in a way that job incumbents could easily relate to them. Positional competencies, comprising intrinsic and extrinsic elements are also considered and benchmark levels are established for positional competencies.

In Chapter 10 *Cloud Cover Forecasting for Solar Plants Using Sequential Sky Images*, Nath, Chaudhary, Sakhet, Gupta, Alex, and Srivastava investigate image forecasting using both classical time series models and deep learning methods. The study finds VAR model is effective in predicting pixel values based on nearby pixels, and it outperforms models that incorporate weather variables as exogenous, and a deep learning-based LSTM model in image population forecasting. Additionally, an unsupervised technique utilizing the K-means algorithm with RB-ratios is proposed that is proven superior to grayscale image analysis.

In Chapter 11 *Enhancing Crop Health Monitoring: The Use of Convolutional Neural Networks for Early Blight and Late Blight Identification*, Roy, Rahul, and Roy examine the utilization of Convolutional Neural Networks (CNNs) to detect late and early blight diseases in potato crops via leaf optical images. The work assesses CNN’s capability in accurately identifying various blight syndromes and healthy potato leaves. Employing a dataset of potato leaf images, the study applies deep learning techniques for model training. Evaluation of the model's performance focuses on its accuracy in distinguishing between different blight diseases and healthy foliage.

In Chapter 12 *Use of Artificial Intelligence for Detection of Data Mining Errors in Real-Time Customer Databases*, Bhattacharjee presents two error detection techniques, namely Isolation Forest and one-class SVM, each applied to measure distinct types of errors. The work compares the performance of these techniques using sixteen parameters. Additionally, it explores two application areas for each method.

Although the chapters in this volume do not delve into the fundamental theories of the discussed topics, they provide concise discussions of relevant principles and basics to ensure comprehensive coverage. Therefore, while some background knowledge in data science may be beneficial, readers are not required to possess advanced expertise in these areas. We anticipate that this volume will serve as a valuable resource for individuals interested in exploring various applications of data science across diverse domains. The primary audience for this book includes advanced postgraduate and doctoral students in finance, econometrics, management, data science, computer science, and information technology. Additionally, faculty members at graduate schools and universities, as well as data science practitioners in the industry, are likely to find the content highly beneficial.

We extend our heartfelt gratitude to all the contributors who have dedicated their time and expertise to the chapters included in this volume. Their invaluable contributions have been instrumental in making this project a success. We would also like to express our sincere appreciation to Cambridge Scholars Publishing for granting us the opportunity to publish our work under their esteemed publishing banner. Special recognition is owed to Adam Rummens and Sophie Edminson from Cambridge Scholars Publishing for their unwavering patience, collaboration, and support throughout the extensive publishing process. Additionally, we would like to acknowledge the indispensable support and cooperation received from our esteemed faculty colleagues at Praxis Business School, Kolkata, India. Their cooperation has been indispensable, and without their involvement, the publication of this volume would not have been achievable. We extend our heartfelt thanks to every one of them for their invaluable contributions and support.

Jaydip Sen & Sayantani Roy Choudhury

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

A COMPARATIVE STUDY OF PORTFOLIO OPTIMIZATION METHODS FOR THE INDIAN STOCK MARKET

JAYDIP SEN, ARUP DASGUPTA AND SAYANTANI ROY CHOUDHURY

Introduction

The design of optimized portfolios has remained a research topic of broad and intense interest among researchers of quantitative and statistical finance for a long time. An optimum portfolio allocates the weights to a set of capital assets in a way that optimizes the return and risk of those assets. Markowitz in his seminal work proposed the *mean-variance optimization* approach which is based on the mean and covariance matrix of returns (Markowitz, 1952). The *mean-variance portfolio* (MVP) design works on an algorithm, known as the *critical line algorithm* (CLA). The CLA algorithm, despite the elegance of its theoretical framework, has some major limitations. One of the major problems is the adverse effects of the estimation errors in its expected returns and covariance matrix on the performance of the portfolio. Since it is extremely challenging to accurately estimate the expected returns of an asset from its historical prices, it is a popular practice to use either a minimum variance portfolio or an optimized risk portfolio with the maximum Sharpe ratio as better proxies for the expected returns. However, due to the inherent complexity, several factors have been used to explain the expected returns.

The *hierarchical risk parity* (HRP) algorithm attempts to address three major shortcomings of quadratic optimization methods which are particularly relevant to the CLA used in the MVP approach to portfolio design (de Prado, 2016). These problems are instability, concentration, and underperformance. Unlike the CLA, the HRP algorithm does not require the covariance matrix of return values to be invertible. Hence, the HRP portfolios can deliver good results even if the covariance matrices are ill-degenerated or singular, which is an impossibility for a quadratic optimizer like CLA. Interestingly, even though MVP's objective is to minimize the variance, HRP is proven to have a lower probability of yielding lower out-of-sample variance than MVP. Given the fact that future returns cannot be forecasted with sufficient accuracy, many researchers have proposed risk-based asset allocation using the covariance matrix of the returns. However, this approach brings in another problem of instability that arises because the quadratic programming methods require the inversion of a covariance matrix whose all eigenvalues must be positive (Baily & de Prado, 2012). HRP is a new portfolio design method that addresses the pitfalls of the quadratic optimization-based MVP approach using techniques of graph theory and machine learning (de Prado, 2016). This method exploits the features of the covariance matrix without the requirement of its invertibility.

The HERC portfolio optimization uses an integrated approach to machine learning and a top-down recursive bisection method of the HRP portfolio method (Raffinot, 2018). The proponents of the HERC method identified several shortcomings of the HRP portfolio optimization. The single linkage-based cluster trees constructed in the HRP method usually led to deep and wide trees and suboptimal allocation of weights to the clusters. The HRP algorithm usually involves higher computations. Finally, the recursive bijection approach used in HRP bisects the cluster tree before the weight allocation instead of directly allocating the weights based on the dendrogram of clustering. This makes the computed weights inaccurate. The HERC algorithm avoids these problems using a top-down recursive bisection and a naive risk parity within the clusters.

This chapter presents a comparative study of the three portfolio optimization methods, MVP, HRP, and HERC, on the Indian stock market, particularly focusing on the stocks chosen from 15 sectors listed on the National Stock Exchange (NSE) of India. The top stocks of each cluster are identified based on their free-float market capitalization from the NSE's report published on July 1, 2022 (NSE Website). For each sector, three portfolios are designed on stock prices from July 1, 2019, to June 30, 2022, following three portfolio optimization approaches. The portfolios are tested over the period from July 1, 2022, to June 30, 2023. For evaluation of the performances of the portfolios, three metrics are used (i) cumulative returns, (ii) annual volatilities, and (iii) Sharpe ratios. based on their cumulative returns. For each sector, the portfolios that yield the highest cumulative return, the lowest volatility, and the maximum Sharpe Ratio over the training and the test periods are identified.