Stock Price Prediction using Technical Indicators and LSTM Based Models

Project Report Submitetd
in partial fulfillment of the requirements for the award of the

Bachelor of Technology

by

Animesh Padhy
(B420007)
Virag Jain
(B420063)
Rohit Gupta
(B420041)

Under the supervision of
Dr. Puspanjali Mohapatra

Department of Information Technology
International Institute of Information Technology, Bhubaneswar

May 2024

APPROVAL OF THE VIVA-VOCE BOARD

May 15, 2024

Certified that the report entitled **Stock Price Prediction using Technical Indicators and LSTM Based Models** submitted by **Animesh Padhy** (B420007), **Virag Jain** (B420063), **Rohit Gupta** (B420041) to International Institute of Information Technology Bhubaneswar in partial fulfillment of the requirements for the award of the Bachelor of Technology in Information Technology under the BTech Programme has been accepted by the examiners during the viva-voce examination held today.

(Supervisor)	**(Panel Head)**
(Internal Examiner 1)	**(Internal Examiner 2)**

CERTIFICATE

This is to certify that the report entitled **Stock Price Prediction using Technical Indicators and LSTM Based Models** submitted by **Animesh Padhy** (B420007), **Virag Jain** (B420063), **Rohit Gupta** (B420041) to International Institute of Information Technology Bhubaneswar is a record of bonafide project work under my supervision, and the report is submitted for end-semester evaluation of the BTech Degree project work.

(Supervisor)

DECLARATION

We certify that

1. The work contained in the report has been done by our group under the general supervision of our supervisor.
2. The work has not been submitted to any other Institute for any degree or diploma.
3. We have followed the guidelines provided by the Institute in writing the thesis.
4. We have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.
5. Whenever we have used materials (data, theoretical analysis, and text) from other sources, we have given due credit to them by citing them in the text of the thesis and giving their details in the references.
6. Whenever we have quoted written materials from other sources, we have put them under quotation marks and given due credit to the sources by citing them and giving required details in the references.

Animesh Padhy
B420007

Virag Jain
B420063

Rohit Gupta
B420041

ACKNOWLEDGMENTS

We wish to express our sincere gratitude to numerous individuals who have contributed to shaping this thesis. First and foremost, we extend our heartfelt appreciation to Prof. Pushpanjalwe Mohapatra for her unwavering guidance throughout the journey of this research endeavor. As our supervisor, she consistently encouraged us to remain focused on achieving our objectives. Her insightful observations and constructive feedback played a pivotal role in guiding the overall direction of the research and in propelling us forward to delve deeper into the subject matter. Her vast knowledge and expertise have been an invaluable asset to us.

We are also thankful to all the esteemed professors of the department for their support and contributions. Our sincere thanks go out to all our friends and well-wishers who have provided us with kind words, a listening ear, fresh perspectives, valuable critiques, or their precious time.

Animesh Padhy
Virag Jain
Rohit Gupta

ABSTRACT

The report delves into the challenges of predicting stock prices due to market volatility and non-linear relationships, and proposes a solution approach utilizing a combination of technical indicators and Long Short-Term Memory (LSTM) based models. This study evaluates 15 research articles and multiple github repositories to gauge the efficacy of algorithms in enhancing the prediction precision. The study evaluates the performance of these models using metrics like R-squared and Mean Absolute Percentage Error (MAPE), demonstrating their potential in providing valuable insights for investment strategies and minimizing risks in the financial market1. The report concludes with suggestions for future research directions, including the exploration of ensemble methods, hybrid models, and the incorporation of news sentiment analysis and social media data for a more holistic approach to stock market prediction.

Keywords: Stock Price Prediction, Feature Engineering, ARIMA, Technical Indicators, LSTM, Machine Learning, Deep Learning

List of Tables

List of Figures

Contents

Chapter 1

Introduction

The stock market is inherently volatile, with prices fluctuating frequently due to various factors such as economic indicators, company performance, and investor sentiment. These rapid changes make it difficult to accurately predict future price movements.

External events can significantly impact stock prices. Economic factors like interest rate changes or GDP growth, social factors such as consumer behaviour trends, and political events like elections or policy changes can all influence market movements. Predicting how these events will unfold and their precise impact on stock prices is complex.

Stock price movements often exhibit non-linear relationships with underlying factors. Traditional linear models may struggle to capture the complexity of these relationships, leading to inaccurate predictions.

In response to these perpetual challenges, the integration of Machine Learning (ML) and Deep Learning (DL) techniques has emerged as a promising frontier in the quest for more accurate predictions.

This study delves into the domain of stock market prediction, with a primary focus on leveraging large historic datasets, statistical models and deep learning techniques to enhance forecasting accuracy. The study aims to address the fundamental problem of predicting stock prices with greater precision, empowering investors to optimize their investment strategies while minimizing risks.

Chapter 2

Literature Survey

The study [1] delves into stock market prediction, leveraging machine learning algorithms to analyze historical data and forecast stock prices amidst market volatility. By comparing five machine learning algorithms; Linear Regression, KNN, Support Vector Regression, Decision Tree Regression and LSTM on the data from 12 prominent Indian companies, the study evaluates their performance. Researchers cleaned and split the data into training and testing sets, training models with different algorithms. Performance metrics like Symmetric Mean Absolute Percentage Error and R-squared were employed to assess model accuracy. Results reveal the Long Short-Term Memory algorithm as the top performer, followed by the Support Vector Regression algorithm, offering valuable insights for effective stock price prediction strategies.

The study [2] presents a new method, ELR-ML, for forecasting financial stock markets using historical data from the S&P 500 index. Through data collection from Yahoo Finance and pre-processing, ELR-ML predicts future S&P 500 values based on various factors. The methodology involves data analysis, estimation, decision-making, and performance evaluation using metrics like R-squared, MSE, SD, and RMSE. Utilizing a decade-long dataset, the paper employs descriptive statistics and comparative analysis against a simple linear regression model to demonstrate the accuracy and reliability of ELR-ML. Graphs and tables further illustrate the findings, confirming the efficacy of the proposed approach for stock market prediction.

The study [3] evaluates the performance of seven machine learning algorithms; Decision trees, Random forest, KNN, Naive Bayes, Logistic regression, SVM and ANN in predicting stock market index movements across developed countries. It investigates indices including NYSE 100, NIKKEI 225, FTSE 100, CAC 40, DAX 30, FTSE MIB, and TSX3 using daily historical data from 2012 to 2021. Technical indicators such as moving averages, momentum, stochastic measures, and relative strength index are input variables, with the next day's movement direction as the output. The study splits the dataset into 80 percent for training and 20 percent for testing, assessing prediction accuracy through metrics like accuracy, precision, recall, F1-score, and area under the curve. Results indicate that artificial neural networks perform optimally for NYSE 100, FTSE 100, DAX 30, and FTSE MIB, while logistic regression is effective for NIKKEI 225, CAC 40, and TSX indices. Notably, all three algorithms achieve prediction accuracies exceeding 70 percent.

The study [4] analyzes ensemble methods for stock market prediction across four countries using data from the Ghana, Johannesburg, New York, and Bombay Stock Exchanges from January 2012 to December 2018. Techniques such as bagging, boosting, stacking, and blending with decision trees, SVMs, and neural networks are evaluated. With wavelet transform for noise reduction and min-max normalization, 25 ensemble models are constructed. Performance assessment involves 12 metrics like accuracy, RMSE, and F1-score. Decision tree ensembles, especially boosting and bagging, achieve high accuracy, followed by neural network ensembles with perfect accuracy over the NYSE dataset. Neural network ensembles outperform decision tree and SVM ensembles, with stacking and blending showing promise for classification and regression tasks.

The study [5] introduces a robust forecasting model for Indian fintech stock prices, focusing on Policybazaar, One 97 Communications Paytm Ltd., and Niyogin Ltd. Utilizing high-frequency one-minute closing prices and the Random Forest algorithm, the study follows a quantitative approach encompassing data collection, preprocessing, model development, and evaluation. Data is sourced from reputable financial databases, covering the period from October 1, 2022, to September 30, 2023, with 70employs metrics like MAE, MSE, RMSE, and R2, with out-of-sample forecasting accuracy assessed through a rolling window approach. Results showcase the Random Forest model's exceptional predictive capabilities, offering low MSE, RMSE, and MAE values alongside high R2 values. Graphical comparisons between actual and predicted stock prices affirm the model's effectiveness, positioning it as a reliable tool for forecasting fintech stock prices in India and providing valuable insights for stakeholders in the sector.

The study [6] explores how well Deep Learning techniques work, particularly RNN and LSTM models, in predicting financial stock market prices. It endeavours to refine prediction accuracy by analyzing diverse datasets and adjusting epochs using RNN and LSTM methodologies. The model operates through a sequential pipeline, collecting user inputs such as preferred stocks, investment amount, risk tolerance, and duration, followed by machine learning algorithms analyzing historical data to forecast future stock trends, aiming to provide optimal investment solutions. Data is sourced from prominent stock exchanges like the National Stock Exchange (NSE), with datasets from leading companies like TCS, Microsoft, Infosys, and TATA obtained from Yahoo Finance. Experimental analysis conducted on a system with 4 GB RAM, an i3 processor, and Jupyter Notebook for Python implementations shows that the LSTM model consistently produces accurate results, affirming its reliability in predicting financial stock market prices.

The study [7] employs LSTM models to analyze stock price trends and predict future behaviour across diverse stocks like Verizon, Netflix, Salesforce, and Amazon. It utilizes deep learning techniques to examine moving averages, correlations between stock prices, and forecast trends. The LSTM model, with two layers, handles information overload during training and enhances accuracy. The analysis focuses on key stock elements such as close, open, high, low, and volume. Using Yahoo Finance data for training and testing, the system demonstrates Root Mean Square Error (RMSE) and prediction trends, showing close alignment between trained, predicted, and validated data. Over time, predictions converge towards reality, with a Train Score of 12.8953 RMSE indicating enhanced accuracy with larger datasets during training.

The study [8] evaluates 12 research papers on machine learning and deep learning techniques for stock price prediction across various sectors. Papers were selected via Google Scholar based on keywords related to machine learning, deep learning, and stock prediction. Findings reveal that artificial neural networks (ANN) and random forest (RF) yield promising results for certain stocks but require extensive data. Long short-term memory (LSTM) models prove effective for others but face overfitting concerns. Apple Inc.'s stock benefits from LSTM combined with sentiment analysis, while XGBoost models are sensitive to outliers. LSTM demonstrates superior performance for stocks from the Dhaka Stock Exchange compared to other methods like XGBoost, linear regression (LR), moving average (MA), and the last value algorithm, though MA is prone to abnormalities.

The study [9] focuses on applying machine learning and deep learning techniques to forecast the stock price of Reliance Industries Limited. It utilizes the ARIMA model for a 2-year forecast and employs Random Forest and LSTM models to predict the next day's stock price. Data preprocessing is conducted using Python libraries, and ML automates model development. The dataset, sourced from Yahoo Finance, consists of 6391 records spanning 1996 to 2021. Both ARIMA and Random Forest models produced accurate results, while LSTM also demonstrated reliability, suggesting promising avenues for stock price prediction.

The study [10] investigates the influence of social media and financial news on stock market prediction accuracy, emphasizing feature selection and spam tweet reduction for improved analysis. Data is sourced from Twitter for social media, Business Insider for financial news, and Yahoo Finance for stock data. Preprocessing involves tokenization and cleaning of data. The Random Forest (RF) classifier emerges as consistently accurate, achieving 80.53% accuracy with social media and 75.16% with financial news. RF's ensemble nature and adaptability to diverse features make it suitable for multi-class problems. Furthermore, employing deep learning, notably a neural network (MLP) with 3 hidden layers, enhances prediction accuracy by up to 7.3%, underscoring the significance of network architecture optimization.

The study [11] investigates stock market prediction through machine learning, focusing on ANN and CNN models and emphasizing their importance in financial decision-making and economic indicators like GDP. Employing deep learning algorithms, including backpropagation for ANN and 2-D histogram processing for CNN, the study trains and tests models with NSE stock market data from April 2008 to April 2018, with a case study during the COVID-19 pandemic from November 2019 to August 2020. Results show the CNN model's superiority, achieving 98.92% accuracy compared to ANN's 97.66%, with the CNN model maintaining 91accuracy even during extreme market fluctuations.

The study [12] aims to improve stock market trend prediction using machine learning and deep learning algorithms, comparing their performance with historical data from the Tehran stock exchange. It evaluates various models, including decision trees, support vector classifiers, and recurrent neural networks, using two input approaches: continuous and binary data. Utilizing ten years of historical data from four stock market groups, the methodology involves implementing nine machine learning methods and two deep learning algorithms with specific parameter adjustments for optimal results. The dataset spans from November 2009 to November 2019, comprising open, close, high, and low values for each trading day, used to compute technical indicators for both approaches. Performance metrics such as F1-Score, Accuracy, and ROC-AUC indicate that deep learning methods, particularly RNN and LSTM, outperform other models, with the binary data approach significantly enhancing prediction performance across all models. These findings underscore the efficacy of deep learning algorithms in forecasting stock market trends.

The study [13] underscores the evolution of stock market prediction methodologies, emphasising the role of machine learning in leveraging non-traditional data sources like social media for improved accuracy. Through a systematic literature review process, the study analyzes findings from the last decade, providing valuable insights into the direction of research in stock market prediction. The methodology involved keyword searches across academic databases, followed by the analysis and synthesis of collected literature to categorize relevant studies. The dataset comprised both market and textual data, including historical price-related numerical data and sentiment analysis from various sources. Performance metrics varied across studies, with accuracies ranging from 50% to 90%, highlighting SVM's popularity while recent trends favour ANN and DNN for their accuracy and speed. The integration of market and textual data further enhanced prediction accuracies, addressing evolving challenges in stock market prediction systems.

The study [14] explores the Impact of Input Window Length on the performance of a predictive system for stock price movements. Using daily data for the S&P 500 Stocks, the system computes multiple technical indicators, including SMA, EMA, ATR, ADMIn, CCI, ROC, RSI, William's %R, Stochastic %K, and %D. Various machine learning algorithms, such as SVM, ANN, and kNN are employed to forecast future directions of stock price movements, with model performance evaluated using prediction accuracy, winning rate, return per trade, and Sharpe ratio.

The study [15] introduces a decision support system aimed at predicting stock market movements by integrating machine learning algorithms and technical indicators derived from historical stock data. Using daily data spanning the last five years for the NIFTY Index, Reliance, and TCS Stocks, the system computes six technical indicators, including Force Index, Williams%R, and Relative Strength Index. Various machine learning algorithms, such as K-Nearest Neighbors and Deep

Neural Networks, are employed to forecast price changes, with model performance evaluated using accuracy and kappa values. The system's comprehensive approach seeks to provide accurate buy-or-sell signals and confidence levels, thereby assisting investors in making well-informed investment decisions within the dynamic stock market landscape.

Chapter 3

Problem statement and Solution Approach

Developing an reliable stock prediction model using a combination of technical indicators and Long Short-Term Memory based deep learning models. The goal is to leverage historical stock price data and various technical indicators to forecast future stock prices. The project aims to address challenges such as data preprocessing, feature engineering, model training, and evaluation to create a robust and effective stock prediction system.

3.1 Dataset

3.1.1 Raw data

The dataset utilized in this study originates from Yahoo Finance and comprises historical price data for two major stock market indices: the NIFTY 50 (N̂SEI) and the BSE SENSEX (B̂SESN). This dataset offers a valuable resource for analyzing and understanding the dynamics of these prominent indices over an extensive period.

For the NIFTY 50 index (N̂SEI), the data spans from September 17, 2007, to May 13, 2024, covering a significant timeframe of nearly seventeen years. This extensive temporal range enables to capture of long-term trends, seasonal fluctuations, and other patterns that may influence market behavior over time.

Similarly, the dataset for the BSE SENSEX index ($\hat{B}$SESN) provides an even broader perspective, encompassing data from July 1, 1997, to May 13, 2024. With over twenty-six years of historical data, this dataset offers insights into the evolution of one of the oldest and most widely followed stock market indices in India. The datasets include various attributes for each trading day, including Open, High, Low, and Close prices, as well as Volume. These attributes capture essential aspects of daily trading activity, reflecting the price movements and liquidity of the respective indices. Each observation in the dataset represents a single trading day, with the 'Date' serving as the index.

	Open	High	Low	Close	Volume	Dividends	Stock Splits
Date							
1997-07-01 00:00:00+05:30	4263.109863	4301.770020	4247.660156	4300.859863	0	0.0	0.0
1997-07-02 00:00:00+05:30	4302.959961	4395.310059	4295.399902	4333.899902	0	0.0	0.0
1997-07-03 00:00:00+05:30	4335.790039	4393.290039	4299.970215	4323.459961	0	0.0	0.0
1997-07-04 00:00:00+05:30	4332.700195	4347.589844	4300.580078	4323.819824	0	0.0	0.0
1997-07-07 00:00:00+05:30	4326.810059	4391.009766	4289.490234	4291.450195	0	0.0	0.0
...	...	...	...	...	...	...	...
2024-05-07 00:00:00+05:30	73973.296875	74026.796875	73259.257812	73511.851562	7900	0.0	0.0
2024-05-08 00:00:00+05:30	73225.000000	73684.929688	73073.921875	73466.390625	6500	0.0	0.0
2024-05-09 00:00:00+05:30	73499.492188	73499.492188	72334.179688	72404.171875	9200	0.0	0.0
2024-05-10 00:00:00+05:30	72475.453125	72946.539062	72366.289062	72664.468750	11000	0.0	0.0
2024-05-13 00:00:00+05:30	72476.648438	72863.562500	71866.007812	72776.132812	13900	0.0	0.0

6616 rows × 7 columns

Figure 3.1: Sensex raw data from July 1997

3.1.2 Data Preprocessing

In the preliminary stages of data preprocessing, a critical step involves the removal of initial segments of the dataset, commonly referred to as burn-in or warm-up periods. This practice is pivotal for ensuring the quality and reliability of subsequent analyses and predictive models, particularly in the context of financial data. Discarding these initial segments ensures that the subsequent analyses and models are based on more consistent and representative data, enhancing their accuracy and reliability in capturing underlying patterns and dynamics in the financial markets. Moreover, allowing machine learning models to initialize with a stable subset of data, mitigates the risk of overfitting and enables them to adapt more effectively to the evolving market conditions over time.

The datasets exhibit anomalies in the form of outliers and zero values within the Volume column, indicating potential errors or missing data. Understanding that the Volume metric reflects the number of individuals engaged in trading activities rather than stock prices underscores its significance in market analysis. Moreover, technical indicators like the Force Index rely heavily on Volume, implying that inaccuracies in this attribute could compromise the reliability of associated analyses. Notably, instances where Volume registers zero result in a corresponding zero value for the Force Index, further emphasizing the need for meticulous data handling. To mitigate these issues, one approach involves filtering the dataset to exclude entries where Volume equals zero, thereby refining the dataset for more robust analysis and prediction tasks.

3.1.3 Exploratory Data Analysis

Univariate analysis plays a crucial role in understanding the characteristics and distributions of individual variables within the dataset. To gain insights into the data and identify potential outliers, box plots are utilized. These plots provide visual representations of the distribution of each variable, highlighting the presence of any extreme values or anomalies that may require further investigation. Additionally, density plots are employed to visualize the distribution of data values for each variable. This step is particularly important as it helps ensure that the data distribution remains consistent after preprocessing or outlier treatment.

In the bivariate analysis, the focus shifts to examining the relationship between two variables within the dataset. Specifically, line plots visualize the closing price for dates, providing a clear depiction of how the closing price fluctuates over time. This visual representation enables analysts to identify trends, patterns, and potential seasonality in the data, facilitating a deeper understanding of the underlying dynamics driving price movements.

Additionally, the investigated correlation between the closing prices at different time points, specifically between the closing price at time t (y(t)) and the closing price at time t+5 (y(t+5)) visualises the degree of linear association between these two variables. A high positive correlation suggests a strong positive relationship, indicating that changes in the closing price at time t are

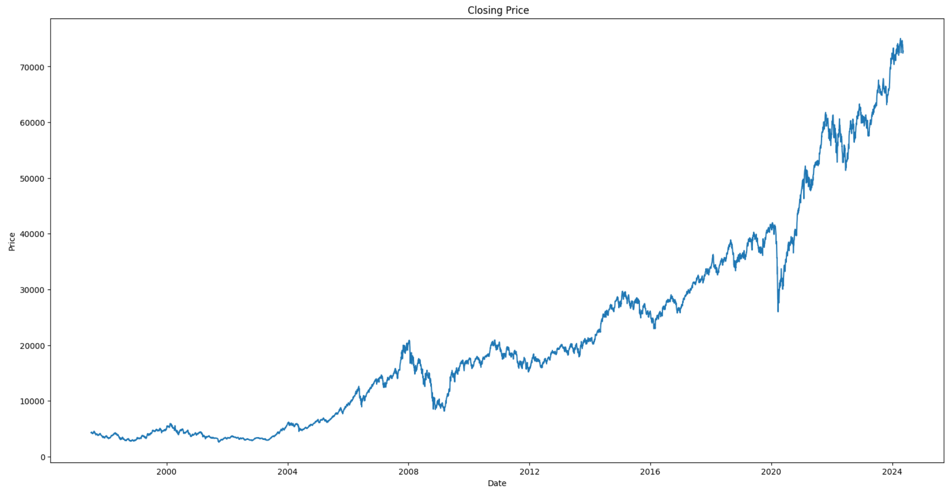

Figure 3.2: Closing Prices for Sensex

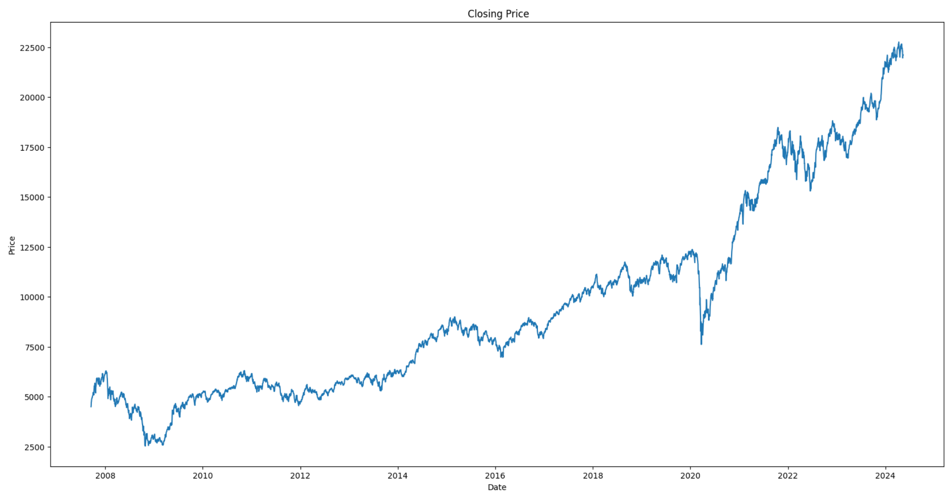

Figure 3.3: Closing Prices for Nifty

closely related to changes in the closing price five time periods later. Conversely, a negative correlation implies an inverse relationship, while a correlation close to zero suggests little to no linear relationship between the two variables.

Multivariate analysis is conducted through the generation of a correlation matrix heatmap. This heatmap visually represents the pairwise correlations between different features within the dataset. By depicting the strength and direction of linear relationships between variables, the correlation matrix heatmap offers valuable insights into the interdependencies among various attributes. Specifically, it enables the identification of potential patterns, associations, or redundancies within

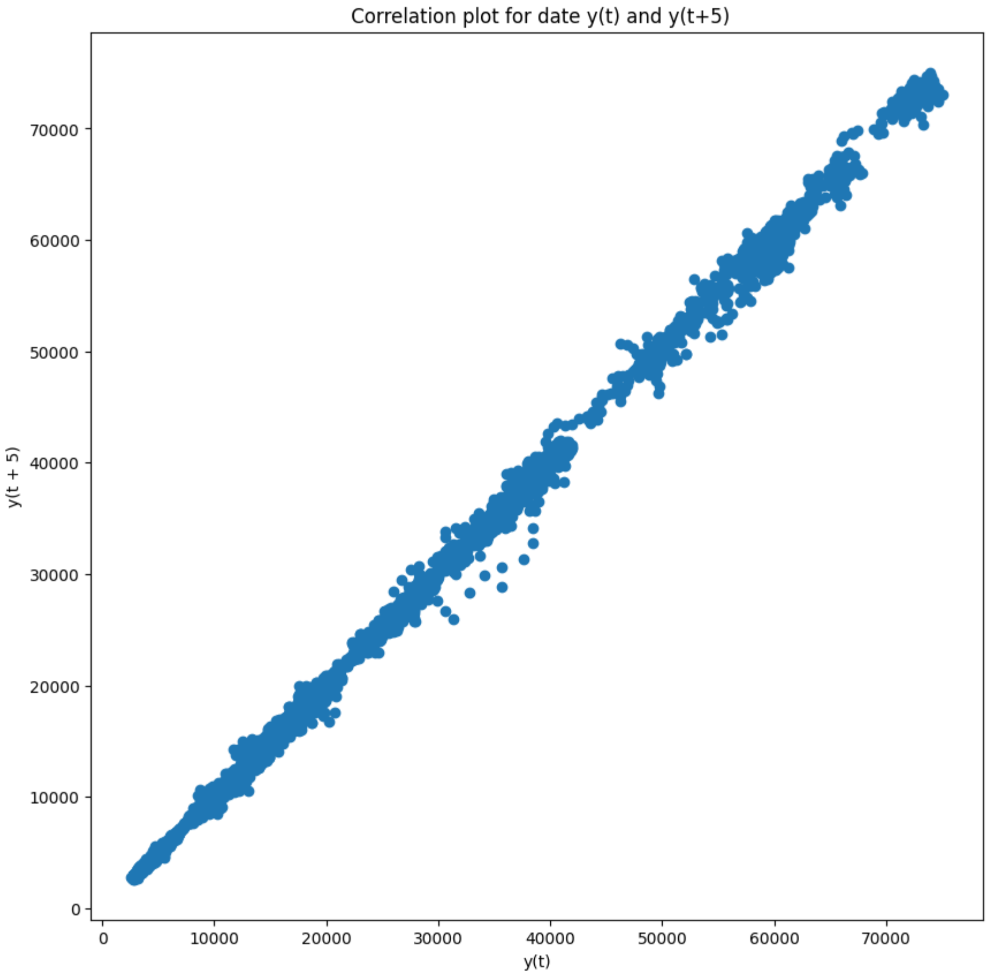

Figure 3.4: Correlation plot for date y(t) and y(t+5)

the data, aiding in feature selection, model development, and interpretation of results.

3.1.4 Data Labelling

The dataset is partitioned into training and testing sets, with the initial 80for training and the latest 20% for testing. This partitioning ensures that the model learns from historical data while being evaluated on unseen data. Moreover, reserving the latest portion for testing enables the evaluation of the model's performance on unseen data, simulating real-world scenarios where the model predicts future prices based on current information. The daily closing prices serve as the target variable for prediction. Training on historical data helps the model understand market patterns while testing on unseen data evaluates its predictive performance. This approach ensures

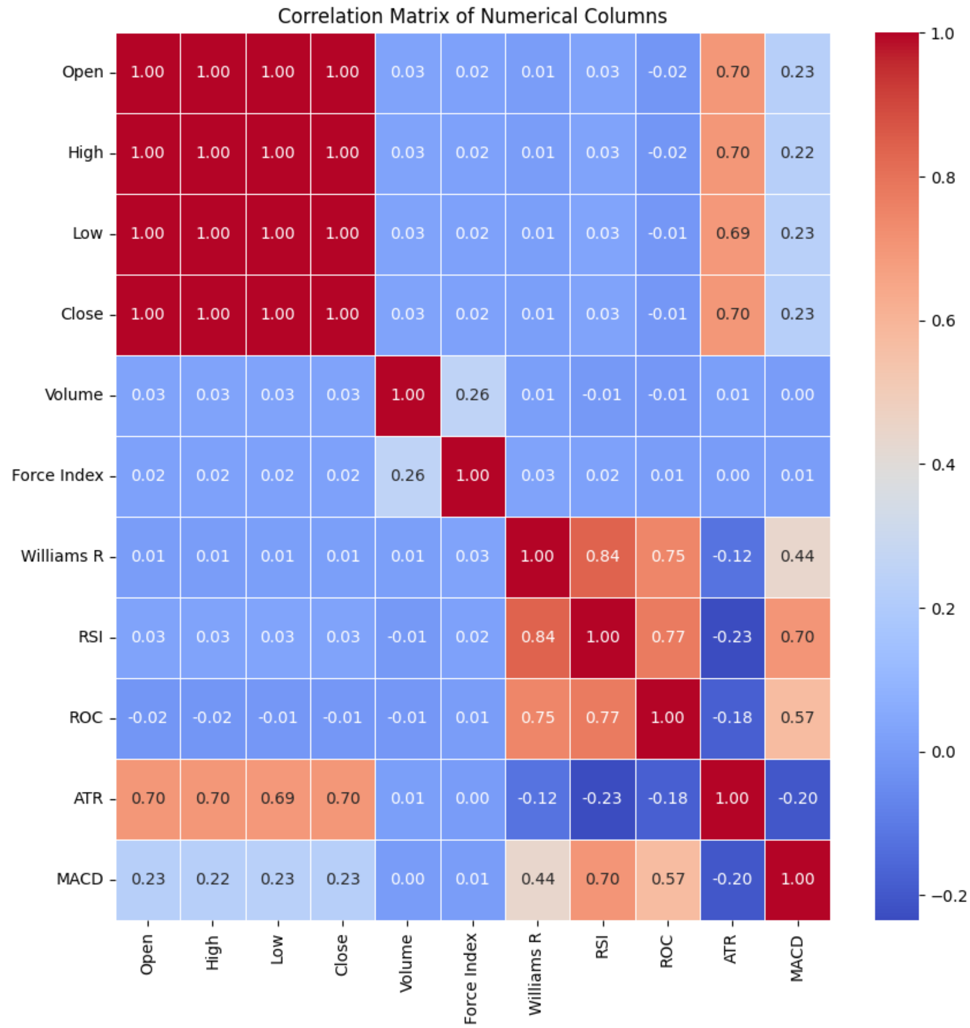

Figure 3.5: Correlation Matrix of Numerical Columns

the model's effectiveness in forecasting future stock prices by learning from past observations and being evaluated on recent data.

3.2 Feature Engineering

Technical indicators are mathematical calculations based on historical price, volume, or open interest data of a financial asset. These indicators are used by traders and analysts to gain insights into the current and future direction of price movements. They help traders make informed decisions about buying, selling, or holding an asset.

Figure 3.6: Train-Test Split for Sensex

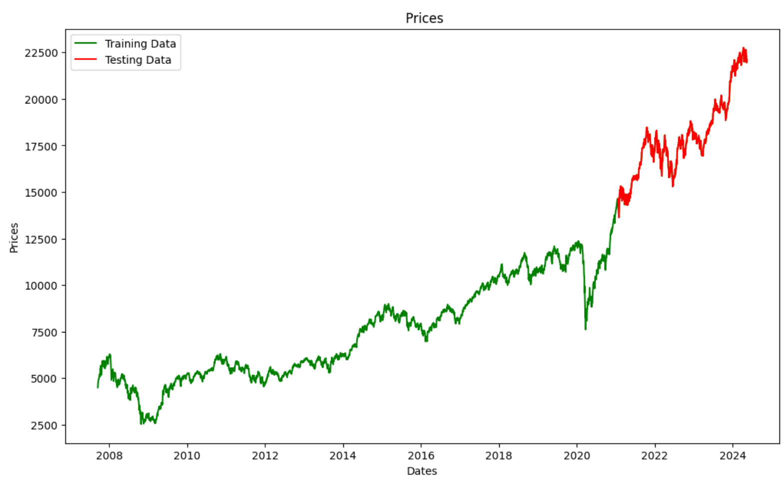

Figure 3.7: Train-Test Split for Nifty

By feeding technical indicators' historical values into a Machine Learning model along with past price data, the model can learn patterns and relationships that may contribute to making more accurate predictions of future closing prices.

3.2.1 Force Index

The Force Index is an oscillator that measures the strength of price movements, combining the price action and volume into a single value. It helps traders identify the strength of a price movement based on trading volume. A rising Force Index suggests strong buying pressure, while a falling Force Index indicates selling pressure.

$$\text{Force Index} = \text{Volume} \times (\text{Close} - \text{Previous Close}) \tag{3.1}$$

3.2.2 Williams Percent Range

Williams %R is a momentum oscillator that measures overbought or oversold levels based on closing prices relative to the high-low range over a specific period (13 days in our case). Values above -20 indicate overbought conditions, while values below -80 suggest oversold conditions. Traders use it to identify potential reversal points.

$$\%R = \frac{(\text{HighestHigh} - \text{Close})}{(\text{HighestHigh} - \text{LowestLow})} \times (-100) \tag{3.2}$$

3.2.3 Relative Strength Index

RSI is a momentum oscillator that measures the speed and change of price movements, indicating overbought or oversold conditions. RSI values above 70 indicate overbought conditions, while values below 30 suggest oversold conditions. It helps traders assess the strength of a trend and potential reversal points.

$$\text{RSI} = 100 - \left(\frac{100}{1 + \text{RS}}\right) \tag{3.3}$$

RS = Average Gain / Average Loss

3.2.4 Rate of Change

ROC measures the percentage change in price over a specific period, indicating the momentum of price movements. Positive ROC values suggest upward momentum, while negative values indicate downward momentum. Traders use it to confirm trends and assess potential entry or exit points.

$$\text{ROC} = \left(\frac{\text{Close} - \text{Close n periods ago}}{\text{Close n periods ago}}\right) \times 100 \tag{3.4}$$

3.2.5 Average True Range

ATR measures market volatility by calculating the average range between high and low prices over a specific period. A higher ATR suggests higher volatility, while a lower ATR indicates lower volatility. Traders use it to set stop-loss levels and determine position sizes based on market volatility.

$$\text{ATR} = \frac{1}{N} \times \sum_{i=1}^{N} (\text{High}_i - \text{Low}_i) \tag{3.5}$$

3.2.6 Moving Average Convergence Divergence

MACD is a trend-following momentum indicator that shows the relationship between two moving averages of an asset's price. Traders use MACD to identify bullish and bearish signals. When the MACD line crosses above the signal line, it indicates a bullish signal, and vice versa for a bearish signal.

$$\text{MACD} = \text{EMA}_{12} - \text{EMA}_{26} \tag{3.6}$$

EMA stands for Exponential Moving Average. It is a type of moving average that gives more weight to recent prices, making it more responsive to short-term price movements compared to simple moving averages.

3.3 Applied algorithms

3.3.1 AutoRegressive Integrated Moving Average

ARIMA is a type of time series analysis and forecasting model that's used in statistics and econometrics to understand and predict patterns in data over time. It consists of following 3 components.

i. AutoRegressive (AR): This part of the model captures the relationship between an observation and a certain number of lagged observations (previous time steps).

ii. Integrated (I): This part of the model deals with making the time series stationary by differencing the observations. Stationarity means that the statistical properties of the time series, like its mean and variance, remain constant over time.

iii. Moving Average (MA): This part of the model uses the dependency between an observation and a residual error from a moving average model applied to lagged observations. It helps in capturing the short-term fluctuations in the time series.

ARIMA models are expressed as ARIMA(p, d, q), where p is the number of lag observations included in the model (order of the AutoRegressive part), d is the number of times the raw observations are differenced (order of differencing) and q is the size of the moving average window (order of the Moving Average part).

Stock prices typically exhibit time-dependent patterns, where today's price is influenced by previous prices. ARIMA models are designed to capture these dependencies by incorporating lagged values into the model. The Moving Average component of ARIMA helps in capturing short-term fluctuations in the stock prices. This is useful for predicting day-to-day changes in prices.

3.3.2 Long Short-Term Memory

LSTM is a type of recurrent neural network (RNN) architecture designed to capture long-term dependencies in sequential data. It addresses one of the key limitations of traditional RNNs, which struggle to learn from and remember long sequences of information due to the vanishing gradient problem. There are two main parts of a LSTM model; Memory cells and Gates.

The core idea behind LSTM is the use of special memory cells that can store information for long periods. These cells maintain a state that can be updated, read, and written to selectively. This ability to retain information over long sequences is what distinguishes LSTMs from standard RNNs.

LSTMs use gates to control the flow of information into and out of the memory cells. These gates are implemented as neural network layers that regulate the information flow based on the input data and the current state of the LSTM. There are 4 types of Gates.

i. Forget Gate: Decides which information to keep and which to forget using a sigmoid activation function. A value close to 1 means "keep," while close to 0 means "forget."

ii. Input Gate: Determines the new information to be added to the cell state. It calculates a candidate vector using the tanh activation function, which creates values between -1 and 1.

iii. Update Cell State: Combines the decisions from the forget gate and the new candidate values to update the cell state.

iv. Output Gate: Produces the output based on the updated cell state.

LSTMs are trained using backpropagation through time (BPTT). During training, the network adjusts its parameters (weights and biases) to minimize the difference between predicted and actual outputs, using optimization algorithms like gradient descent. The use of LSTMs in stock price prediction projects is motivated by their ability to handle sequential data, capture long-term dependencies, and learn complex patterns and relationships.These characteristics allow LSTMs to mske accurate predictions in financial markets.

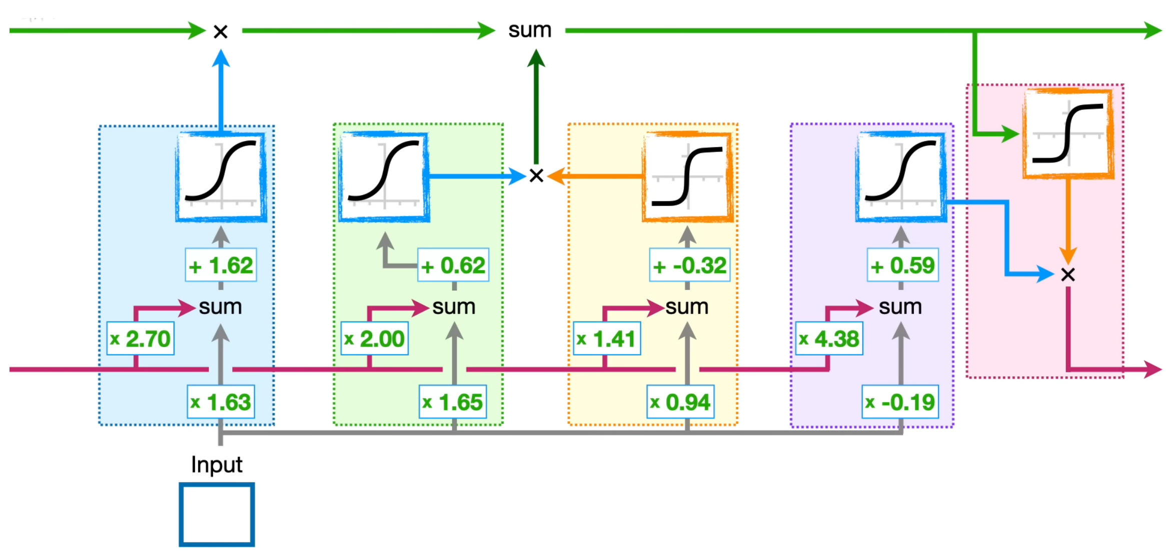

Figure 3.8: LSTM Architecture
[16]

Chapter 4

Results and Discussions

4.1 Performance Metrics

R2 and MAPE are used for evalutating the performances of ARIMA and LSTM models predictions. R-squared assesses how well the independent variables explain the dependent variable's variability in a regression model, while MAPE measures the accuracy of predictions in forecasting or modeling by calculating the average percentage difference between actual and predicted values.

Model	**R2**	**MAPE**
ARIMA (Nifty)	0.993	0.006
ARIMA (Sensex)	0.997	0.008
LSTM (Nifty)	0.985	0.009
LSTM (Sensex)	0.992	0.011

Table 4.1: Comparison of R2 and MAPE for ARIMA and LSTM models

4.1.1 R-squared

R-squared is a statistical measure that represents the proportion of the variance in the dependent variable that is predictable from the independent variables in a regression model. It tells us how well the independent variables in a regression model explain the variability of the dependent variable.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{4.1}$$

R-squared values range from 0 to 1, where 0 indicates that the model does not explain any variability in the dependent variable and 1 indicates that the model perfectly explains the variability in the dependent variable. R2 simply indicates the strength of the relationship between the independent and dependent variables in the model.

4.1.2 Mean Absolute Percentage Error

MAPE is a measure of the accuracy of a forecasting or predictive model. It calculates the average percentage difference between actual and predicted values.

$$\text{MAPE}(y, \hat{y}) = \frac{100\%}{n} \sum_{t=1}^{n} \left| \frac{e_t}{y_t} \right| \tag{4.2}$$

A lower MAPE indicates a more accurate model, as it signifies that the predictions are closer to the actual values on average.

4.2 Performance Visualization

The graphs display the ARIMA and LSTM model predictions for Sensex prices, showing how well the forecasts match actual price movements over time. These visualizations help assess the model's accuracy and guide decisions on trading strategies and forecasting adjustments.

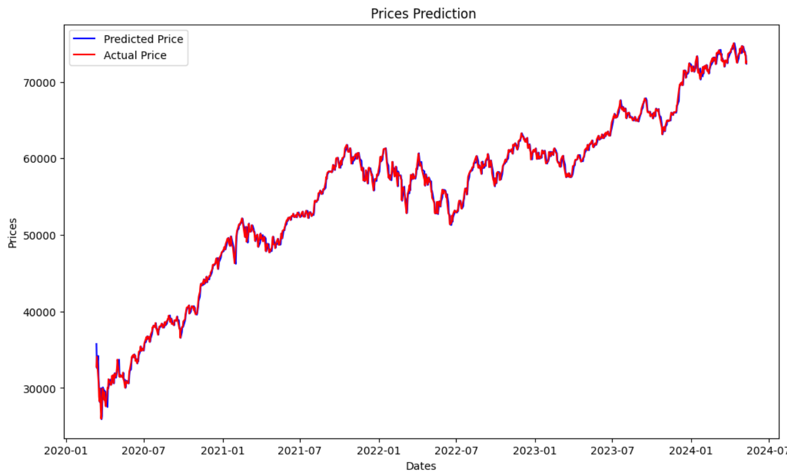

Figure 4.1: Sensex Stock Price Prediction using ARIMA

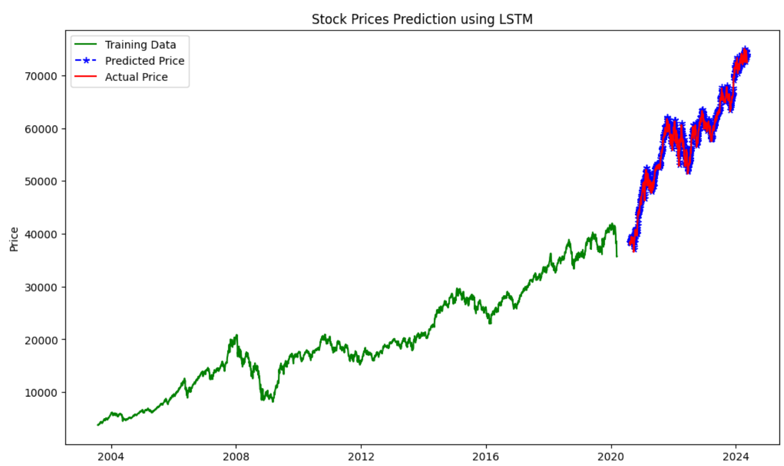

Figure 4.2: Sensex Stock Price Prediction using LSTM

The graphs display the ARIMA and LSTM model predictions for Nifty prices, highlighting the alignment between forecasts and actual price trends over time. These visualizations are instrumental in evaluating the model's precision and informing decisions regarding trading strategies and refining forecasting techniques.

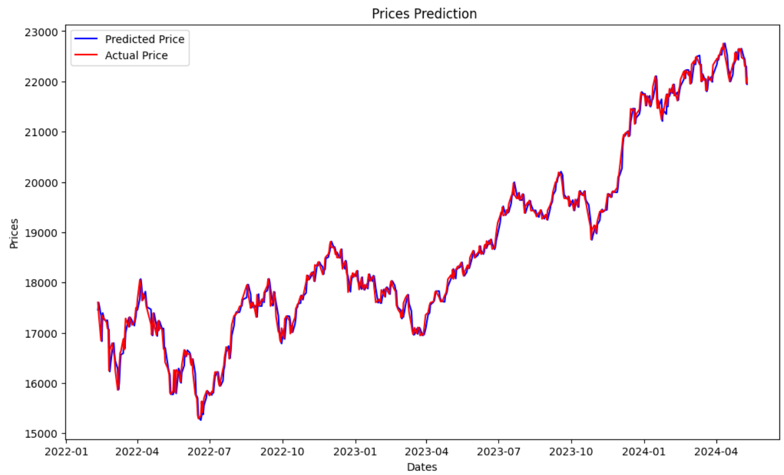

Figure 4.3: Nifty Stock Price Prediction using ARIMA

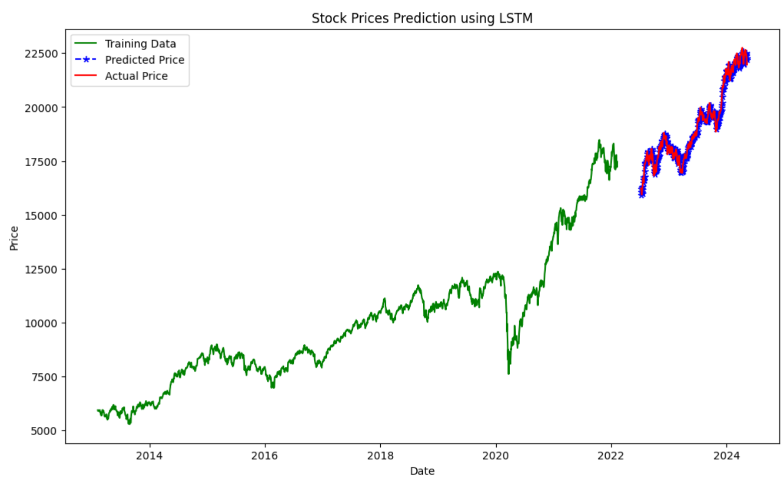

Figure 4.4: Nifty Stock Price Prediction using LSTM

4.3 Performance and Loss Evaluation

1. The Sensex LSTM model reached a minimum validation loss of 0.00019147878629155457 at Epoch 96.

2. In comparison, the Nifty LSTM model achieved a lower minimum validation loss of 0.00014730831026099622 at Epoch 100.

These results underscore the potential of LSTM models in forecasting stock market movements, particularly for benchmark indices like Sensex and Nifty. The models' capacity to minimize validation loss suggests their utility in providing valuable insights into market trends and supporting decision-making processes within the financial domain.

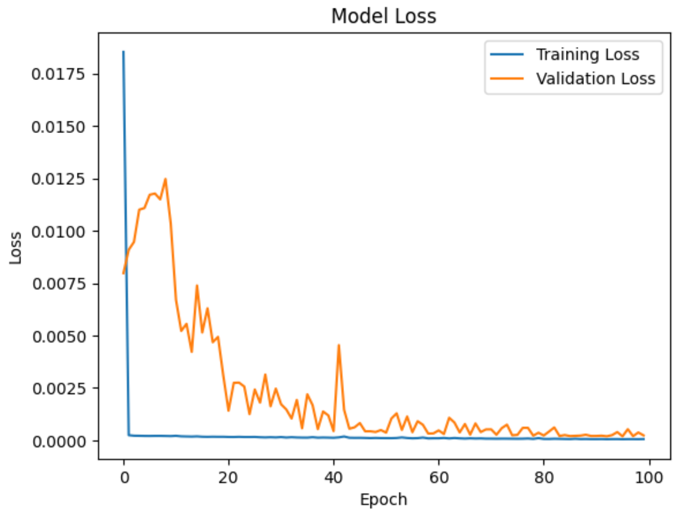

Figure 4.5: LSTM Model Loss - Sensex

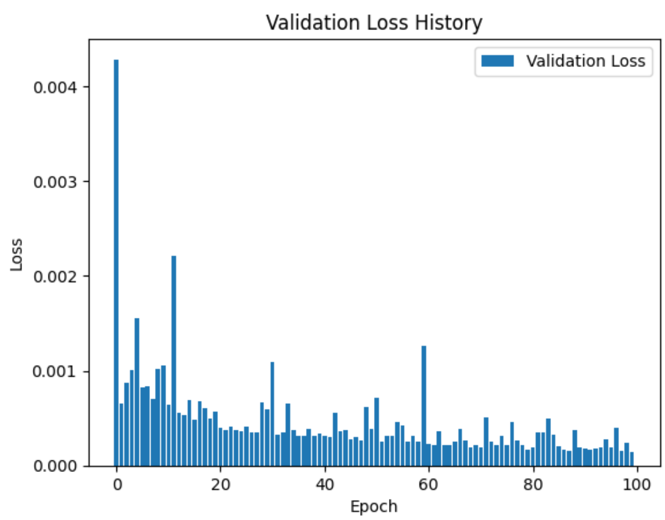

Figure 4.6: LSTM Validation Loss - Nifty

Chapter 5

Conclusion and Future Scope

While this research delves into ARIMA and LSTM models, there's an array of other potential options worth exploring. Ensemble methods like Boosting, Recurrent Neural Networks, hybrid models, and optimization approaches tailored to financial time series data offer promising avenues for our further investigation.

It's crucial to acknowledge that technical indicators alone may not suffice for decision-making in the volatile stock market. Market conditions can change rapidly, necessitating a broader perspective that includes fundamental factors such as company financials and economic news. Incorporating news sentiment analysis, social media data, and alternative datasets could provide a more comprehensive understanding of market dynamics.

Looking ahead, there's potential in framing market price prediction as a classification problem, wherein the objective is to determine whether the price will rise, fall, or remain unchanged compared to the previous day. This approach could offer actionable insights for users, guiding them on whether to hold, buy, or sell stocks on a given day.

Bibliography

[1] Malti Bansal, Apoorva Goyal, and Apoorva Choudhary. Stock market prediction with high accuracy using machine learning techniques. *Procedia Computer Science*, 215:247–265, 2022.

[2] J Margaret Sangeetha and K Joy Alfia. Financial stock market forecast using evaluated linear regression based machine learning technique. *Measurement: Sensors*, 31:100950, 2024.

[3] Nazif Ayyildiz and Omer Iskenderoglu. How effective is machine learning in stock market predictions? *Heliyon*, 10(2), 2024.

[4] Isaac Kofi Nti, Adebayo Felix Adekoya, and Benjamin Asubam Weyori. A comprehensive evaluation of ensemble learning for stock-market prediction. *Journal of Big Data*, 7(1):20, 2020.

[5] Bharat Kumar Meher, Manohar Singh, Ramona Birau, and Abhishek Anand. Forecasting stock prices of fintech companies of india using random forest with high-frequency data. *Meher, BK, Singh, M., Birau, R., & Anand, A.(2023). Forecasting Stock Prices of Fintech Companies of India using Random Forest with High-frequency Data. Journal of Open Innovation: Technology, Market, and Complexity*, 100180, 2023.

[6] K Ritwik Reddy, B Tarun Kumar, V Rohit Ganesh, Polisetty Swetha, and Prakash Kumar Sarangi. Stock market prediction using recurrent neural network. In *2022 IEEE International Conference on Current Development in Engineering and Technology (CCET)*, pages 1–4. IEEE, 2022.

[7] B Shivani and S Govinda Rao. Analysis of a stock exchange and future prediction using lstm. In *2021 4th International Conference on Recent Trends in Computer Science and Technology (ICRTCST)*, pages 48–54. IEEE, 2022.

[8] Adel Ismail Al-Alawi and Yusuf Ahmed Alaali. Stock market prediction using machine learning techniques: Literature review analysis. In *2023 International Conference On Cyber Management And Engineering (CyMaEn)*, pages 153–157. IEEE, 2023.

[9] S Pratheeth and Vishnu Prasad. Stock price prediction using machine learning and deep learning. In *2021 IEEE Mysore Sub Section International Conference (MysuruCon)*, pages 660–664. IEEE, 2021.

[10] Wasiat Khan, Mustansar Ali Ghazanfar, Muhammad Awais Azam, Amin Karami, Khaled H Alyoubi, and Ahmed S Alfakeeh. Stock market prediction using machine learning classifiers and social media, news. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–24, 2022.

[11] Somenath Mukherjee, Bikash Sadhukhan, Nairita Sarkar, Debajyoti Roy, and Soumil De. Stock market prediction using deep learning algorithms. *CAAI Transactions on Intelligence Technology*, 8(1):82–94, 2023.

[12] Mojtaba Nabipour, Pooyan Nayyeri, Hamed Jabani, S Shahab, and Amir Mosavi. Predicting stock market trends using machine learning and deep learning algorithms via continuous and binary data; a comparative analysis. *Ieee Access*, 8:150199–150212, 2020.

[13] Nusrat Rouf, Majid Bashir Malik, Tasleem Arif, Sparsh Sharma, Saurabh Singh, Satyabrata Aich, and Hee-Cheol Kim. Stock market prediction using machine learning techniques: a decade survey on methodologies, recent developments, and future directions. *Electronics*, 10(21):2717, 2021.

[14] Yauheniya Shynkevich, T Martin McGinnity, Sonya A Coleman, Ammar Belatreche, and Yuhua Li. Forecasting price movements using technical indicators: Investigating the impact of varying input window length. *Neurocomputing*, 264:71–88, 2017.

[15] Vimal S Ramanan, A Sheik Abdullah, et al. An efficient decision support system for stock market prediction using technical indicators as features by machine learning approaches. In *2023 4th International Conference for Emerging Technology (INCET)*, pages 1–6. IEEE, 2023.

[16] Josh Starmer. Long Short-Term Memory (LSTM), Clearly Explained, 2023.