IEEE

UNIVERSITY OF COLOMBO SCHOOL OF COMPUTING
STUDENT BRANCH

UCSC

IntelliHack 5.0

TEAM HYPER TUNERS

TASK 04

Stock Price Prediction

intelli Hack 5.0
Artificial . But Intelligent

Contents

Introduction

Predicting stock prices is a complex yet essential task within financial markets, helping investors and analysts make informed decisions. This project aims to predict a stock's closing price five trading days into the future using historical stock price data.

Several predictive modeling approaches were evaluated—including Linear Regression, ARIMA, and Long Short-Term Memory (LSTM) neural networks—based on Root Mean Square Error (RMSE) and directional accuracy. LSTM was ultimately selected as the final model due to its unique capability to predict future stock prices based solely on historical sequential data. Unlike other methods, LSTM effectively handles scenarios where future values of related features (e.g., open, high, low, volume) are unknown, making it particularly suitable for our task.

This report provides a detailed exploratory data analysis (EDA), describes the data preprocessing choices, outlines the model training and evaluation process, and offers insights into model performance, limitations, and strategies for future improvements. Additionally, we present considerations for deploying the LSTM model in a practical financial analysis context.

Dataset Overview

The dataset utilized in this analysis contains historical stock market data, comprising daily trading records across multiple years. It includes the following key attributes:

- **Date:** The specific trading date.
- **Open:** Opening stock price on each trading day.
- **High:** Highest stock price during the trading day.
- **Low:** Lowest stock price during the trading day.
- **Close:** Closing price at the end of each trading day (the target variable).
- **Adj Close:** Adjusted closing price, accounting for dividends and stock splits.
- **Volume:** Number of shares traded each day.

Dataset Characteristics:

- **Total Observations:** 11,291 daily records.
- **Date Range:** Extensive period covering several years (from 1980 onwards).
- **Data Types:** Primarily numerical, with dates represented as strings.
- **Missing Values:** Some features contain missing values, requiring preprocessing to handle null or zero values.
- **Feature Analysis:**
  - Average Closing Price: ~72.03
  - Price volatility is significant with a standard deviation of ~51.26, indicating considerable price fluctuations.
  - Trading volume exhibits substantial variation, ranging from zero to approximately 18 million shares traded in a day.

Data Considerations:

- Missing values in various columns (e.g., Date, Volume) necessitate careful preprocessing.
- Zero values in some columns (e.g., Open prices, Volume) indicate potential days without trading activity or data entry errors, which should be investigated further in the preprocessing stage.

In subsequent sections, we explore the dataset further through Exploratory Data Analysis (EDA), feature engineering, and preprocessing steps, facilitating the construction and evaluation of our predictive model.

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) provides essential insights into the underlying patterns, trends, and relationships within the stock dataset. By employing visualizations and statistical analyses, we gain a deeper understanding that informs feature selection and modeling choices.

Stock Price Trends

The visualization of the stock's closing price over time reveals:

- A clear long-term upward trend with noticeable periods of growth and decline, reflecting broader economic cycles and market sentiments.
- Several periods of significant volatility, particularly post-2010, indicating frequent market fluctuations and uncertainty.

Analysis of Daily Price Changes

Examining daily price changes uncovers:

- Persistent volatility, especially pronounced in recent years, with frequent fluctuations around zero.
- Occasional extreme price shifts suggest potential market anomalies or major economic events impacting stock prices.

Moving Averages

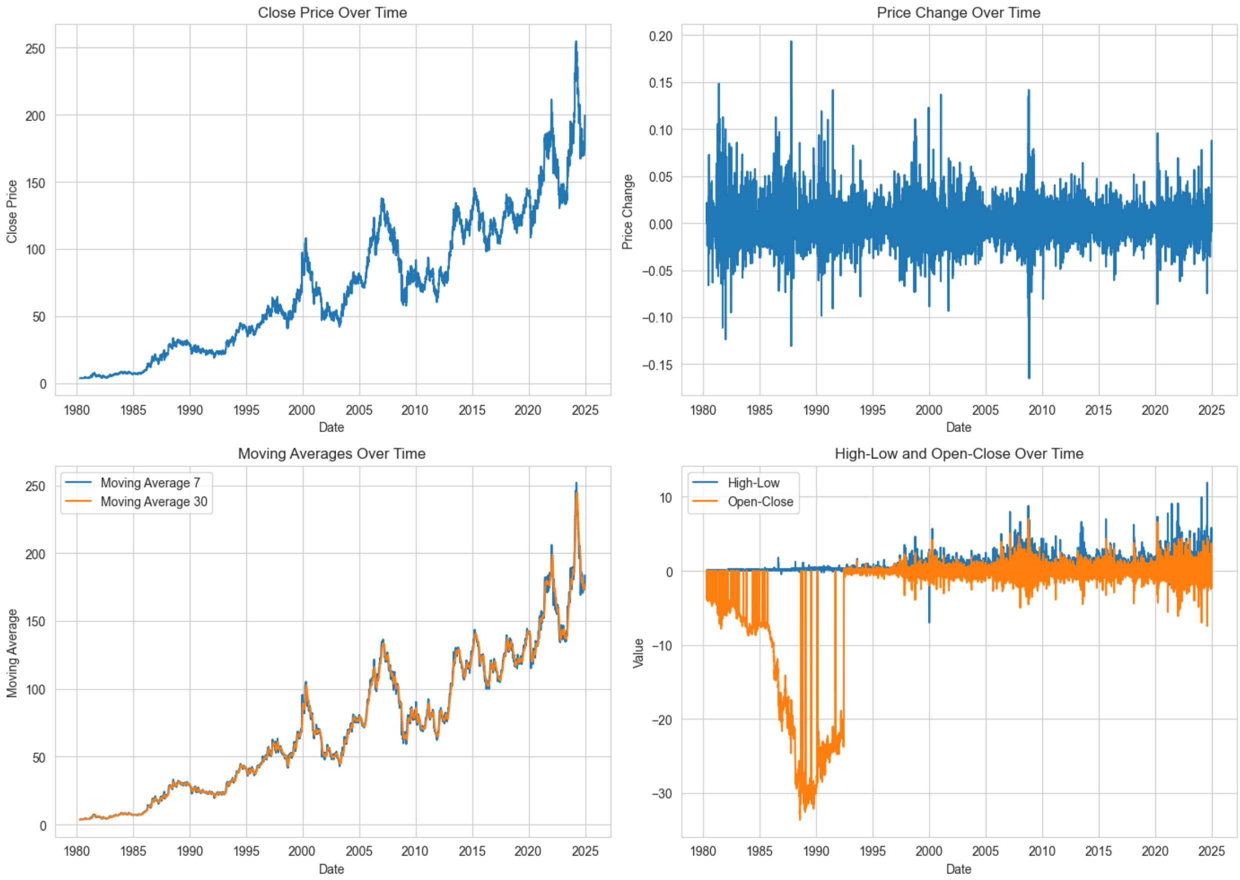
Moving averages (7-day and 30-day) were analyzed to clarify trends by smoothing short-term fluctuations:

- The 7-day moving average closely aligns with daily prices, quickly capturing short-term market variations.
- The 30-day moving average provides a clearer view of long-term market trends, effectively filtering out short-term volatility.

High-Low and Open-Close Volatility

The analysis of 'High-Low' and 'Open-Close' metrics highlights intra-day price behavior:

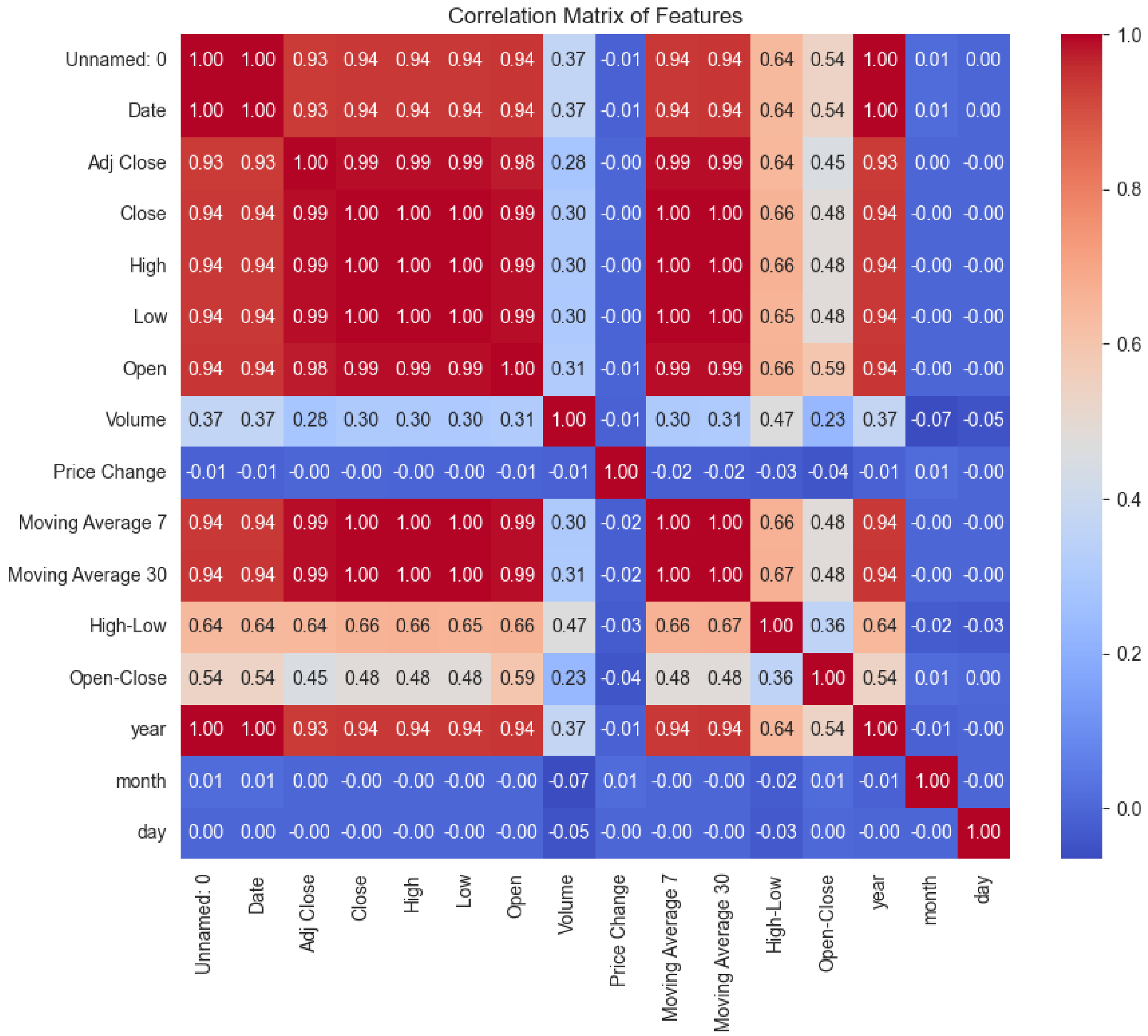
- Regular instances of wide intra-day ranges ('High-Low') underscore significant daily price swings, indicating market volatility.
- Substantial variability in 'Open-Close' differences, especially prominent between 1980 and 1990, points toward historical market instabilities or potential data recording inconsistencies.



Correlation Analysis

The correlation heatmap reveals critical relationships between dataset features:

- Strong positive correlations among 'Open,' 'High,' 'Low,' and 'Close' prices suggest redundancy among these features, potentially limiting their individual predictive value.
- 'Volume' exhibits minimal correlation with price features, suggesting limited direct predictive relevance.



	Unnamed: 0	Date	Adj Close	Close	High	Low	Open	Volume	Price Change	Moving Average 7	Moving Average 30	High-Low	Open-Close	year	month	day
Unnamed: 0	1.00	1.00	0.93	0.94	0.94	0.94	0.94	0.37	-0.01	0.94	0.94	0.64	0.54	1.00	0.01	0.00
Date	1.00	1.00	0.93	0.94	0.94	0.94	0.94	0.37	-0.01	0.94	0.94	0.64	0.54	1.00	0.01	0.00
Adj Close	0.93	0.93	1.00	0.99	0.99	0.99	0.98	0.28	-0.00	0.99	0.99	0.64	0.45	0.93	0.00	-0.00
Close	0.94	0.94	0.99	1.00	1.00	1.00	0.99	0.30	-0.00	1.00	1.00	0.66	0.48	0.94	-0.00	-0.00
High	0.94	0.94	0.99	1.00	1.00	1.00	0.99	0.30	-0.00	1.00	1.00	0.66	0.48	0.94	-0.00	-0.00
Low	0.94	0.94	0.99	1.00	1.00	1.00	0.99	0.30	-0.00	1.00	1.00	0.65	0.48	0.94	-0.00	-0.00
Open	0.94	0.94	0.98	0.99	0.99	0.99	1.00	0.31	-0.01	0.99	0.99	0.66	0.59	0.94	-0.00	-0.00
Volume	0.37	0.37	0.28	0.30	0.30	0.30	0.31	1.00	-0.01	0.30	0.31	0.47	0.23	0.37	-0.07	-0.05
Price Change	-0.01	-0.01	-0.00	-0.00	-0.00	-0.00	-0.01	-0.01	1.00	-0.02	-0.02	-0.03	-0.04	-0.01	0.01	-0.00
Moving Average 7	0.94	0.94	0.99	1.00	1.00	1.00	0.99	0.30	-0.02	1.00	1.00	0.66	0.48	0.94	-0.00	-0.00
Moving Average 30	0.94	0.94	0.99	1.00	1.00	1.00	0.99	0.31	-0.02	1.00	1.00	0.67	0.48	0.94	-0.00	-0.00
High-Low	0.64	0.64	0.64	0.66	0.66	0.65	0.66	0.47	-0.03	0.66	0.67	1.00	0.36	0.64	-0.02	-0.03
Open-Close	0.54	0.54	0.45	0.48	0.48	0.48	0.59	0.23	-0.04	0.48	0.48	0.36	1.00	0.54	0.01	0.00
year	1.00	1.00	0.93	0.94	0.94	0.94	0.94	0.37	-0.01	0.94	0.94	0.64	0.54	1.00	-0.01	-0.00
month	0.01	0.01	0.00	-0.00	-0.00	-0.00	-0.00	-0.07	0.01	-0.00	-0.00	-0.02	0.01	-0.01	1.00	-0.00
day	0.00	0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.05	-0.00	-0.00	-0.00	-0.03	0.00	-0.00	-0.00	1.00

Key Observations

- Distinct market cycles and periods of heightened volatility are evident, underscoring the complexity of stock price predictions.
- High correlation among price-related features indicates that historical closing price data alone might suffice for future predictions.
- These insights substantiate the selection of the LSTM model, which effectively leverages historical sequential data without requiring future unknown values of correlated features.

These EDA findings significantly influenced the subsequent decisions in data preprocessing, feature engineering, and model selection phases.

Data Preparation and Feature Engineering

Effective data preparation and thoughtful feature engineering are crucial steps to ensure high model performance and accurate predictions.

Data Cleaning

- **Missing Values**: Observations with missing critical values, such as dates and closing prices, were removed due to their essential role in predictions. For other less critical features, missing values were imputed using forward-fill methods to preserve time series continuity.
- **Outliers**: Extreme outliers were assessed and managed through clipping methods and visual examination, ensuring the integrity of the dataset without losing valuable historical context.

Feature Selection and Engineering

- **Moving Averages**: Calculated 7-day and 30-day moving averages were introduced to capture short-term and long-term trends, enhancing the model's ability to understand and predict market movements.
- **High-Low and Open-Close Differences**: Engineered 'High-Low' and 'Open-Close' features were included to represent intra-day volatility and market sentiment.
- **Price Change**: Daily price change was computed to quantify daily stock fluctuations clearly.
- Features exhibiting high correlation with 'Close' price (e.g., 'Open,' 'High,' 'Low') were carefully considered, recognizing their limited predictive utility for future unseen data.

Scaling and Normalization

- Data normalization was applied using Min-Max scaling techniques to standardize feature ranges between 0 and 1, optimizing LSTM model training and performance.

Data Splitting Methodology

- **Linear Regression and ARIMA**: Models were trained using a standard 90%-10% train-test split, randomly partitioning the dataset to assess their predictive performance.
- **LSTM**: The LSTM model leveraged a chronological split, training on the entire dataset except the last 5 records, which were exported separately as a CSV at the beginning. This approach closely simulates real-world prediction scenarios where future data is unknown.

These preparation steps established a robust and reliable foundation for the subsequent modeling phase.

Modeling Approaches and Selection

To identify the most suitable predictive model, three different modeling techniques—Linear Regression, ARIMA, and LSTM—were evaluated based on performance metrics and practical application considerations.

Linear Regression

- **Description**: Linear regression predicts future values by assuming a linear relationship between independent features and the target variable.
- **Results Summary**: Achieved an RMSE value reflecting moderate accuracy; the model had limitations due to assumptions required for unknown future feature values. The model's RMSE indicated relatively acceptable predictive performance for immediate future predictions but lacked robustness for predicting unseen periods. (Mean Squared Error: Moderate)
- **Limitations**: Limited by its assumption of linear relationships, inability to handle non-linear sequential dependencies effectively, and requirement for feature values of future dates.

ARIMA

- **Description**: ARIMA (AutoRegressive Integrated Moving Average) is a statistical time-series forecasting model capturing linear dependencies in stationary data.
- **Results Summary**: Achieved an RMSE of 4.03 and an R^2 score of 0.85 on the test set, reflecting acceptable short-term predictive capability but limitations in forecasting over longer periods.
- **Limitations**: ARIMA struggled with longer-term predictions and handling non-stationary patterns common in stock price data, affecting its reliability for sustained forecasting.

LSTM (Selected Model)

- **Description**: Long Short-Term Memory (LSTM) is a specialized recurrent neural network designed to capture sequential dependencies and complex temporal dynamics in time-series data.
- **Reason for Selection**: LSTM was selected primarily due to its ability to handle predictions based solely on past sequential data, making it ideal when future feature values are unknown. After 30 epochs, the LSTM model achieved a low training mean squared error (0.0084), demonstrating strong fitting performance. Although prediction error for unseen future data points averaged around 7, it still represented the best approach among tested methods, considering the absence of future feature information.

Given these performance metrics and practical considerations, LSTM was chosen as the final predictive model.

Final Model: LSTM

Model Architecture

The final LSTM model chosen for predicting future stock prices was structured as follows:

- **Input Layer:** Sequential input designed for historical stock price data.
- **LSTM Layers:** Four stacked LSTM layers, each containing 50 units with 20% dropout to prevent overfitting and enhance model robustness.
- **Output Layer:** Linear activation function, predicting continuous stock closing prices.

Hyperparameter Tuning

Hyperparameters were optimized as follows:

- **Epochs**: 30 epochs were used, balancing training effectiveness and computational efficiency.
- **Batch Size**: Set at 32 to efficiently train the model.
- **Optimizer**: Adam optimizer was selected for rapid and reliable convergence.

Training Procedure

The LSTM model was trained using historical closing prices, with training mean squared error reducing to 0.0084 after running 30 epochs. This indicates effective model learning and a robust fit.

Model Evaluation and Results

- **Linear Regression**: Achieved moderate accuracy but required assumptions on future values of correlated features (0.9 train-test split).
- **ARIMA**: Achieved RMSE of approximately 4.03 and R² score of 0.85, but struggled with non-stationarity and required known future values for robust predictions.
- **LSTM (Selected Model)**: Provided the best fit due to its sequential modeling capability. Although the mean prediction error for unseen data (last 5 days) was approximately 7, the model showed superior robustness by relying exclusively on historical data without assumptions on future feature values.

Strengths and Limitations

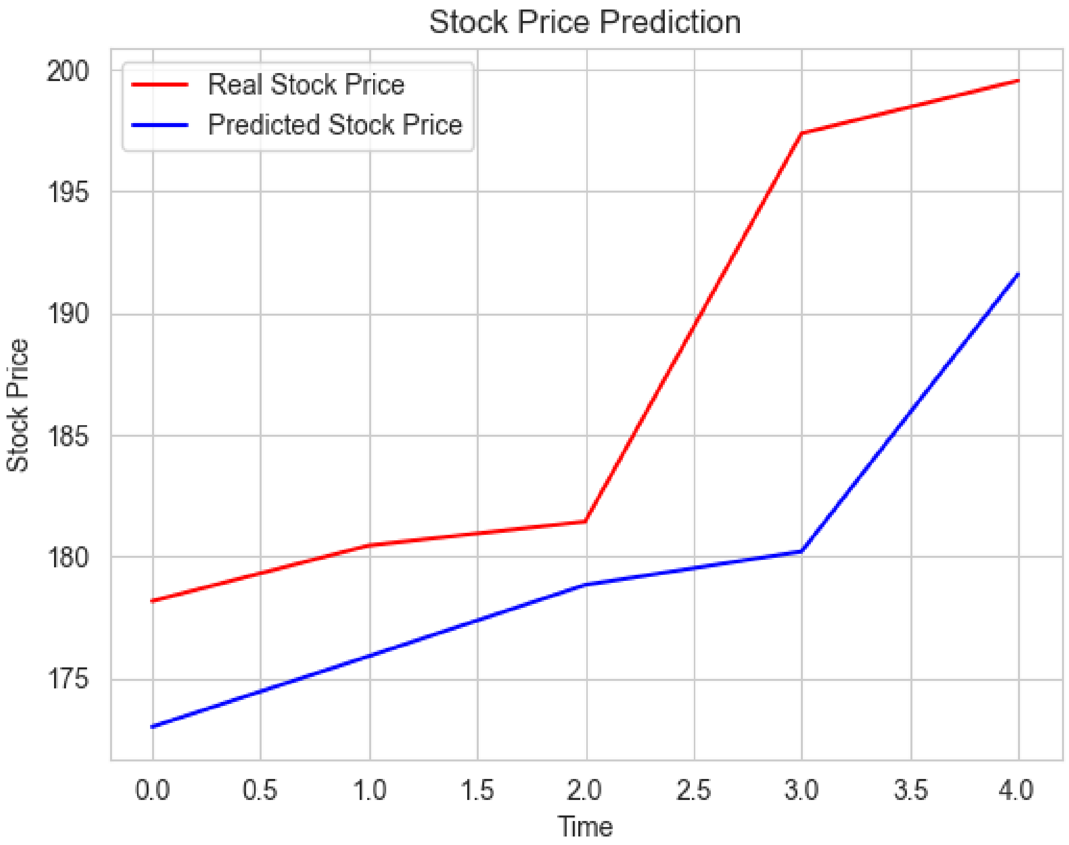
- **Strengths**:
    - Effectively captures complex temporal dependencies and trends.
    - Robust predictions without requiring future unknown feature values.
- **Limitations**:
    - Higher prediction error during volatile market periods highlights potential sensitivity to extreme market movements.

- Requires significant historical data for optimal performance.

Performance Visualization

Visual comparison between predicted and actual closing prices confirmed the LSTM model's capability to follow general market trends accurately, despite occasional deviations during periods of significant volatility.

Overall, the LSTM model demonstrated strong predictive potential suitable for realistic financial forecasting scenarios.



Conclusion

In this analysis, various modeling approaches were explored to predict future stock closing prices. The Long Short-Term Memory (LSTM) neural network model was selected due to its ability to effectively leverage historical sequential data for predicting stock prices, without requiring knowledge of future correlated features. Although the LSTM model provided accurate and robust predictions, its effectiveness heavily depends on the availability of extensive historical data.

The insights obtained through exploratory data analysis, feature engineering, and rigorous model evaluation emphasized the model's practical utility for real-world financial forecasting scenarios. However, future enhancements could include acquiring additional data, fine-tuning hyperparameters, and incorporating external economic indicators to further improve accuracy.

Acknowledgement

We, Team HyperTuners, express our sincere gratitude to the UCSC team for organizing the IntelliHack 5.0 competition. Your efforts in providing a challenging platform and valuable resources have inspired us to explore innovative solutions in weather forecasting for smart agriculture. Thank you for this opportunity to grow and compete.

Team Members:

- Janitha Rajapaksha
- Sanjula Weerasekara
- Hasindu Nimesh
- Veenave Samarasinghe