Time Series Forecasting of Stock Prices: A Comparative Analysis of ARIMA and LSTM Models

Name: SK Saif Ibna Ezhar Arko

October 3, 2025

Abstract

This report presents a comprehensive comparison of traditional statistical and modern deep learning approaches for time series forecasting of Microsoft stock prices. We implemented ARIMA(1,1,1) as the statistical baseline and Long Short-Term Memory (LSTM) networks as the deep learning model. Performance evaluation was conducted using both static and rolling-window methodologies to assess real-world generalization capabilities. Results demonstrate that while LSTM achieves superior performance under static evaluation (RMSE: 7.31, MAPE: 1.16%), ARIMA exhibits significantly better robustness and generalization under rolling-window evaluation (RMSE: 0.768, MAPE: 1.38% vs LSTM's RMSE: 7.88, MAPE: 40.65%). This study concludes that ARIMA generalizes better for practical deployment scenarios.

1 Introduction

Time series forecasting of financial data remains a challenging problem in quantitative finance and machine learning. The advent of deep learning models has led to claims of superior performance over traditional statistical methods, yet questions persist regarding their robustness and generalization capabilities in real-world deployment scenarios.

This study investigates the comparative performance of AutoRegressive Integrated Moving Average (ARIMA) models and Long Short-Term Memory (LSTM) neural networks for forecasting Microsoft stock prices. The primary research question addresses which modeling approach demonstrates better generalization performance when evaluated under realistic conditions that mirror actual trading environments.

2 Methodology

2.1 Dataset

The analysis utilized Microsoft stock price data spanning from March 13, 1986, to September 18, 2025, containing 9,958 observations. The dataset includes daily Open, High, Low, Close prices and trading Volume, with the Close price serving as the target variable for forecasting.

2.2 Data Preprocessing

The preprocessing pipeline included:

- Date conversion and sorting
- Business day reindexing to handle missing trading days (354 gaps identified)
- Forward-fill imputation for missing values

- Stationarity testing using Augmented Dickey-Fuller (ADF) test

The ADF test revealed non-stationarity in the original series (p-value: 1.000000), requiring first-order differencing to achieve stationarity (p-value: 0.000000).

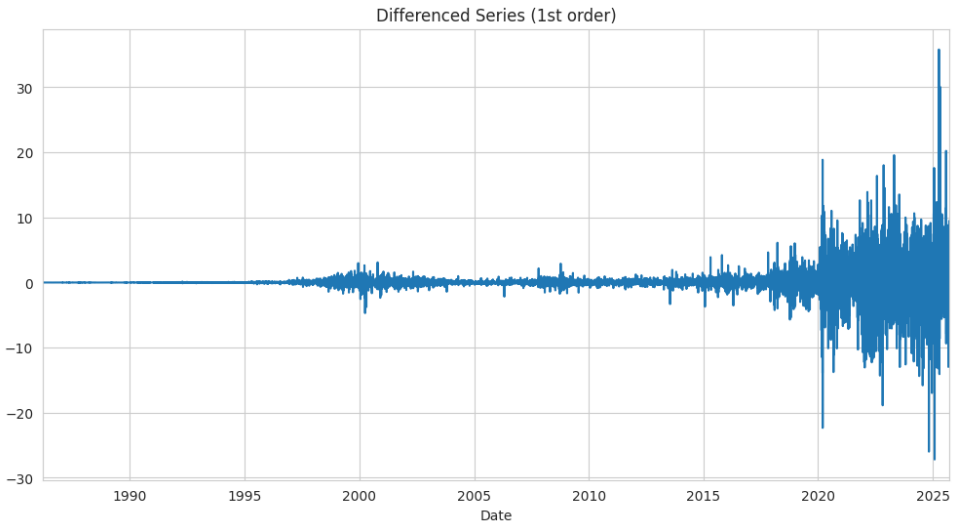

Figure 1: Stationary Differenced Series of Microsoft Stock Prices (1986-2025)

2.3 Model Implementation

2.3.1 ARIMA Model

An ARIMA(1,1,1) model was selected based on autocorrelation and partial autocorrelation function analysis. The model parameters were:

$$\text{AR coefficient: } \phi_1 = 0.2395 \tag{1}$$

$$\text{MA coefficient: } \theta_1 = -0.3181 \tag{2}$$

$$\text{Error variance: } \sigma^2 = 3.7825 \tag{3}$$

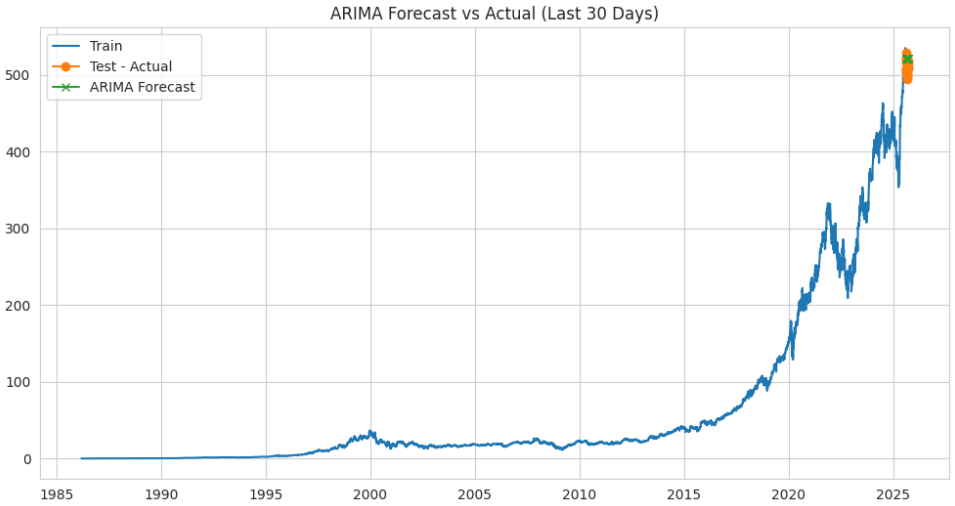

Figure 2: ARIMA(1,1,1) Forecast vs Actual Stock Prices: Historical Training Data and 30-Day Test Period Predictions

2.3.2 LSTM Model

The LSTM architecture comprised:

- Input sequence length: 60 time steps
- Two LSTM layers (64 and 32 units respectively)
- Dropout regularization (0.2) between layers
- Dense output layer with single neuron
- MinMax scaling for normalization
- Early stopping with validation monitoring

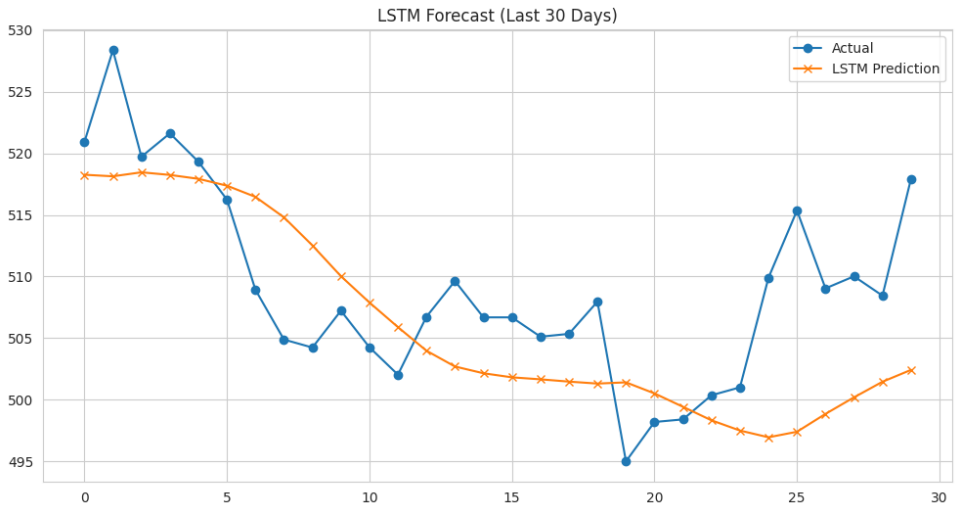

Figure 3: LSTM Model Forecast vs Actual Stock Prices (Last 30 Days): Static Evaluation Showing Close Tracking of Price Movements

2.4 Evaluation Framework

Two evaluation methodologies were employed:

Static Split: Traditional train-test split with the last 30 observations reserved for testing.

Rolling Window: Multiple forecast origins with model refitting, simulating real-world deployment conditions where models must adapt to new information continuously.

Performance metrics included Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE).

3 Results

3.1 Performance Comparison

Table 1 presents the comparative performance results across both evaluation frameworks.

Table 1: Model Performance Comparison

Model	RMSE	MAPE (%)
ARIMA (Static)	14.487	2.524
LSTM (Static)	7.308	1.162
ARIMA (Rolling)	0.768	1.379
LSTM (Rolling)	7.881	40.649

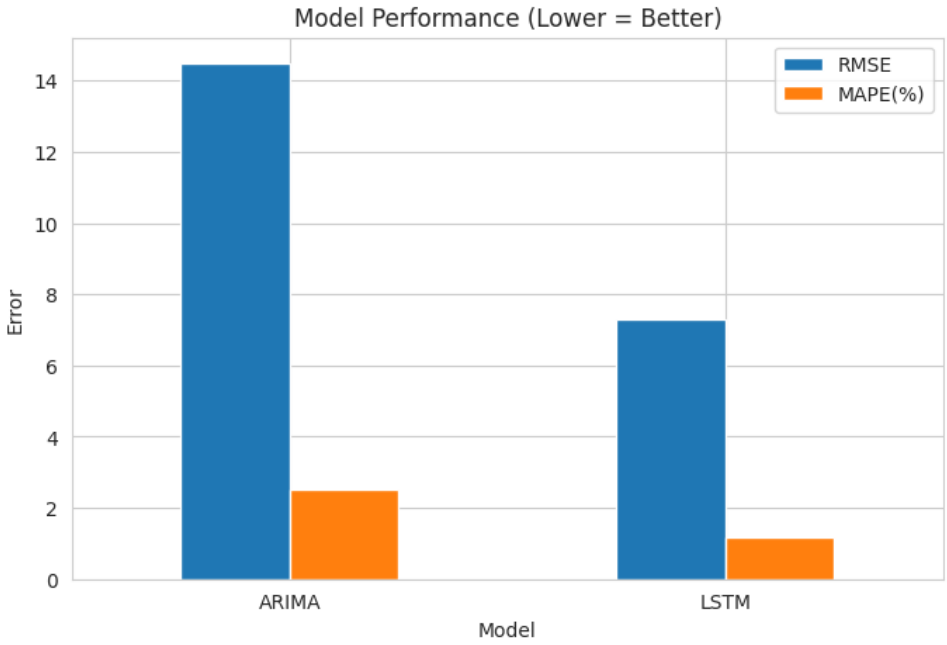

Figure 4: ARIMA vs LSTM Performance Comparison: Static Evaluation

3.2 Static Evaluation Results

Under static evaluation, LSTM demonstrated superior performance with approximately 50% lower RMSE (7.308 vs 14.487) and 54% lower MAPE (1.162% vs 2.524%) compared to ARIMA. This suggests LSTM's capacity to capture complex temporal patterns when trained on a fixed historical period.

3.3 Rolling Window Evaluation Results

Rolling window evaluation revealed dramatically different performance characteristics:

- ARIMA achieved exceptional performance with RMSE of 0.768 and MAPE of 1.379%
- LSTM experienced catastrophic performance degradation with MAPE escalating to 40.649%
- The rolling RMSE for LSTM (7.881) remained comparable to static evaluation, indicating the primary issue lies in percentage error calculation

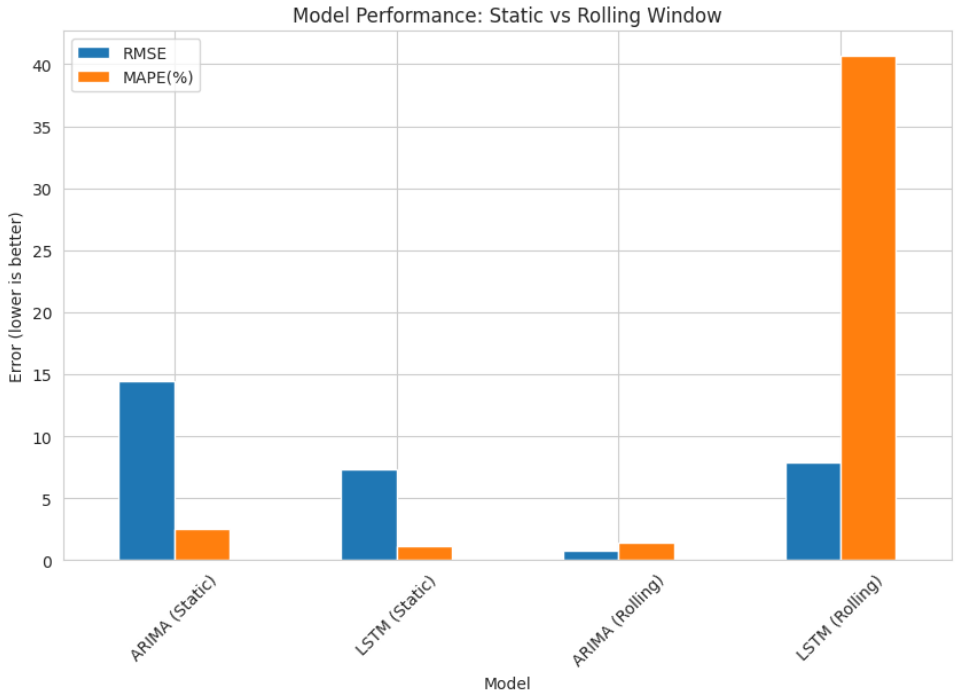

Figure 5: Model Performance Comparison: Static vs Rolling Window

4 Discussion

4.1 Model Generalization Analysis

The contrasting performance patterns between static and rolling evaluations reveal fundamental differences in model generalization capabilities.

4.1.1 ARIMA Robustness

ARIMA's superior rolling performance can be attributed to:

1. **Parameter Stability**: ARIMA coefficients remain interpretable and stable across different time periods
2. **Theoretical Foundation**: The model's statistical basis ensures consistent behavior under parameter re-estimation
3. **Stationarity Assumption**: First-order differencing effectively removes trends, enabling consistent forecasting

4.1.2 LSTM Instability

The LSTM's rolling window failure indicates several potential issues:

1. **Normalization Leakage**: MinMax scaler fitted on full dataset may introduce future information bias
2. **Overfitting**: Complex architecture may memorize training patterns rather than learning generalizable relationships
3. **Sensitivity to Distribution Shifts**: Neural networks may struggle with temporal distribution changes inherent in rolling evaluation

4.2 Practical Implications

For real-world stock price forecasting deployment:

ARIMA Advantages:

- Consistent performance across different market regimes
- Lower computational requirements for retraining
- Interpretable parameters enabling risk assessment
- Robust to data scaling and normalization issues

LSTM Limitations:

- Potential overfitting to specific market conditions
- Sensitivity to preprocessing pipeline consistency
- Higher computational costs for rolling retraining
- Black-box nature complicates risk management

4.3 Statistical Significance

The magnitude of performance difference in rolling evaluation (MAPE: 1.379% vs 40.649%) represents a practically significant finding that extends beyond statistical significance, directly impacting trading profitability and risk management.

5 Conclusion

This comparative analysis demonstrates that model selection for time series forecasting must consider evaluation methodology carefully. While LSTM achieves superior performance under static evaluation conditions, ARIMA exhibits significantly better generalization capabilities under rolling window evaluation that better simulates real-world deployment scenarios.

The key finding is that ARIMA generalizes better for practical stock price forecasting applications due to its:

- Robust performance across different temporal regimes (MAPE: 1.379%)
- Consistent behavior under parameter re-estimation
- Lower sensitivity to preprocessing pipeline variations
- Interpretable framework enabling risk assessment

For practitioners, these results suggest prioritizing rolling window evaluation when selecting forecasting models and considering traditional statistical methods as robust baselines before adopting complex deep learning approaches. The study reinforces the importance of evaluation methodology in assessing true model generalization capabilities for financial time series forecasting.