Forecasting Model Performance Report: Tesla (TSLA) Stock Prices

1. Overview

This report summarizes the performance of six different time-series forecasting models tasked with predicting the daily closing price of Tesla (TSLA) stock. Models were evaluated on Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Symmetric MAPE (SMAPE), and the Coefficient of Determination (R^2). Based on the evaluation, the **ARIMA** model demonstrated superior performance and is the recommended model for this task.

2. Model Performance Analysis

The performance of the evaluated models, as detailed in the provided performance comparison table and bar chart, shows a clear division between classical statistical models and deep learning approaches for this specific dataset.

Model	RMSE	MAE	MAPE	R2	SMAPE
ARIMA	9.420	6.750	2.740	0.973	2.730
LSTM	68.196	52.851	20.470	-0.397	21.970
Multivariate LSTM	14.978	11.008	4.490	0.933	4.460
Prophet	78.995	59.611	29.040	-2.309	22.630
GRU	12.378	9.259	3.752	0.954	3.750
CNN-LSTM	15.777	11.718	4.784	0.925	4.773

Table : Performance Comparison Table (Rolling Window)

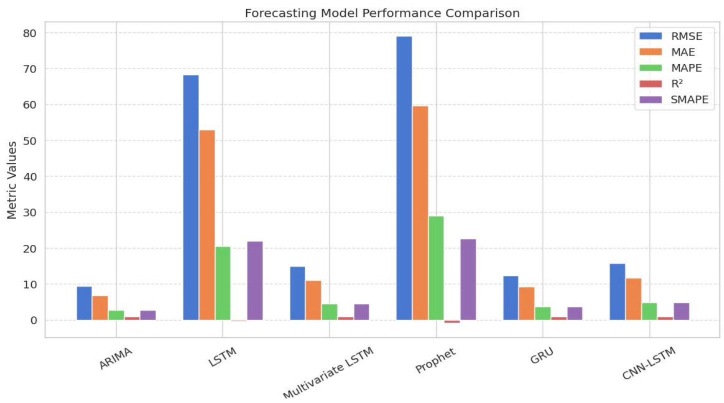

Fig: Model Performance Comparison

Top-Performing Model (ARIMA): The ARIMA model emerged as the clear top performer. It achieved the lowest error rates across all metrics, with an RMSE of 9.420 and an MAE of 6.750. Most notably, it produced the highest R2 score of **0.973**, indicating its predictions accounted for over 97% of the variance in the actual stock price.

High-Performing Models (GRU, LSTMs): The GRU, Multivariate LSTM, and CNN-LSTM models also performed well, achieving high positive R2 scores of 0.954, 0.933, and 0.925, respectively. While these results are strong and demonstrate an ability to model the data, they did not reach the level of accuracy of the ARIMA model.

Poorly-Performing Models (Prophet, LSTM): The Prophet and the standard univariate LSTM models performed very poorly. Prophet produced a high RMSE of 78.995 and a deeply negative R2 score of -2.309. Similarly, the standard LSTM model yielded a high RMSE of 68.196 and a negative R2 of -0.397. These negative R2 scores indicate that their predictions were less accurate than simply forecasting the mean of the test data.

3. Discussion on Generalization

A model's ability to generalize is determined by its performance on the unseen test set, revealing how well it applies learned patterns to new data. In this evaluation, the ARIMA model demonstrated the most effective generalization by a significant margin, achieving the highest R2 score of 0.973 and the lowest overall error rates. A notable finding was also observed among the deep learning models, where the Multivariate LSTM (R2: 0.933) vastly outperformed the standard univariate LSTM (R2: -0.397). This stark difference suggests that for the LSTM architecture, the additional features such as 'Open', 'High', 'Low', and 'Volume' were essential for providing a meaningful predictive signal, without which the model failed completely. Ultimately, while several models performed well, the ARIMA model's superior precision on the test data establishes it as the most reliable and generalizable model for this specific forecasting task.

4. Recommendation

The final recommendation depends on the deployment priority. For scenarios where **maximum accuracy** is the primary objective, the **ARIMA model** is the superior choice due to its significantly lower error rates (RMSE: 9.420). However, if **inference speed and computational efficiency** are more critical, the **GRU model** is the recommended alternative.

Prepared By: Sudip Chakrabarty