Stock Price Forecasting Project Report

May 8, 2025

Prepared by: Grok 3, xAI

Based on the Python Notebook: grok_suggestion.py

Contents

1 Introduction

This report presents a comprehensive analysis of the stock price forecasting project implemented in the Python notebook grok_suggestion.py. The project develops and evaluates forecasting models for stock prices using statistical time series techniques (AR, ARIMA, SARIMAX, SARIMA-GARCH) and machine learning (LSTM), enhanced by real-time data integration and interactive visualizations. The dataset includes historical stock prices from November 23, 2015, to November 20, 2020, extended with real-time data for Apple Inc. (AAPL) from December 31, 2021, to May 8, 2025. This report explains each of the 22 cells in the notebook, incorporating specific results and outputs from the notebook execution.

1.1 Objectives

- Develop and compare multiple forecasting models.
- Integrate real-time data for current predictions.
- Provide interactive visualizations for stakeholders.
- Identify the best-performing model using standardized metrics.

1.2 Dataset

The primary dataset (yahoo_stock.csv) includes:

- Columns: Date, Open, High, Low, Close, Volume, Adjusted Close
- Period: November 23, 2015, to November 20, 2020 (1825 entries)
- Real-time data: AAPL stock from December 31, 2021, to May 8, 2025, via yfinance

1.3 Tools and Libraries

- Data Manipulation: pandas 2.2, numpy 1.23.5
- Visualization: plotly 5.22
- Time Series: statsmodels 0.14, pmdarima 2.0, arch 6.3
- Machine Learning: tensorflow 2.12, keras 2.12
- Real-Time Data: yfinance 0.2.40

2 Methodology and Analysis

The notebook is organized into 22 cells, each addressing a specific task. Below, each cell is analyzed with its purpose, actions, results, and rationale.

2.1 Cell 1: Import Libraries

Purpose: Set up the environment by installing and importing libraries.

Actions:

- Install: pmdarima 2.0.4, arch 7.2.0, tensorflow 2.12, keras 2.12, numpy 1.23.5.

- Import libraries for data manipulation, visualization, time series, and machine learning.
- Suppress warnings for cleaner output.

Results:

- All libraries installed successfully, with dependencies (e.g., joblib 1.5.0, scipy 1.15.2) verified.
- Output: ”Libraries imported successfully.”

Rationale:

- Specific versions ensure compatibility in Google Colab.
- Comprehensive imports cover all tasks.
- Warning suppression improves readability.

2.2 Cell 2: Load and Prepare Data

Purpose: Load and preprocess the stock price dataset.

Actions:

- Load yahoo_stock.csv, parsing dates, or generate dummy data if unavailable.
- Convert Date to datetime, set as index, sort chronologically.
- Select Close price as the time series (ts_data_close).
- Display data info, first five rows, and missing values.

Results:

- Data loaded successfully (1825 entries, November 23, 2015, to November 20, 2020).
- Columns: High, Low, Open, Close, Volume, Adj Close (all non-null).
- First five rows showed stable prices (e.g., Close: 2086.59 to 2090.11).

Rationale:

- Error handling ensures continuity.
- Datetime index enables time series analysis.
- Close price is the standard forecasting target.

2.3 Cell 3: Exploratory Data Analysis (EDA)

Purpose: Explore trends, patterns, and statistical properties.

Actions:

- Compute descriptive statistics.
- Create Plotly plots: closing price, daily returns, volume, moving averages (50-day, 200-day), Bollinger Bands, daily return histogram.

Results:

- Statistics: Mean Close = 2647.86, Std = 407.30, Min = 1829.08, Max = 3626.91.

- Plots showed trends, volatility, and trading patterns.

Rationale:

- Statistics summarize distribution.
- Visualizations identify trends and volatility.
- Technical indicators are standard in finance.

2.4 Cell 4: Time Series Analysis (Stationarity and Decomposition)

Purpose: Analyze stationarity and decompose the time series.

Actions:

- Perform Augmented Dickey-Fuller (ADF) test.
- Decompose series (period=365) into trend, seasonal, and residual components.
- Plot decomposition components.

Results:

- ADF Test: Statistic = -0.8704, p-value = 0.7976 (non-stationary).
- Decomposition showed weak annual seasonality.

Rationale:

- ADF test confirms need for differencing.
- Decomposition guides model selection.

2.5 Cell 5: Seasonality Validation (ACF/PACF)

Purpose: Validate seasonality using autocorrelation plots.

Actions:

- Test periods: weekly (s=5), monthly (s=12), annual business (s=252), annual daily (s=365).
- Plot ACF and PACF for each period.
- Recommend s=12 for SARIMAX.

Results:

- ACF/PACF showed spikes at monthly lags (12, 24, 36).
- Output: ”Choose s=12 for SARIMAX if monthly patterns are evident.”

Rationale:

- ACF/PACF identify significant lags.
- Monthly seasonality is practical.

2.6 Cell 6: AR Model with Out-of-Sample Testing

Purpose: Implement a baseline AR model.

Actions:

- Split data (80% train = 1460, 20% test = 365).
- Scale data with MinMaxScaler.
- Fit AR(2) model, compute in-sample RMSE.
- Forecast out-of-sample, compute RMSE, plot results.

Results:

- In-sample RMSE: 17.2755
- Out-of-sample RMSE: 264.6477
- Plot showed forecast divergence from actuals.

Rationale:

- AR(2) is a simple baseline.
- Scaling improves stability.
- Out-of-sample testing ensures realism.

2.7 Cell 7: ARIMA Model with Out-of-Sample Testing

Purpose: Develop an ARIMA model with differencing.

Actions:

- Fit ARIMA(1,1,1).
- Compute in-sample RMSE, AIC, BIC.
- Forecast out-of-sample, compute RMSE, plot results.

Results:

- In-sample RMSE: 17.2872
- AIC: 12462.6640, BIC: 12478.5205
- Out-of-sample RMSE: 278.4383

Rationale:

- ARIMA captures trends and noise.
- (1,1,1) balances complexity.

2.8 Cell 8: SARIMAX Model with Exogenous Variables

Purpose: Enhance ARIMA with seasonality and volume.

Actions:

- Use volume as exogenous variable.
- Fit SARIMAX(1,1,1)(1,1,0,12).
- Compute metrics, forecast, plot results.

Results:

- In-sample RMSE: 37.7224

- AIC: 12904.0664, BIC: 12930.4075
- Out-of-sample RMSE: 442.4877

Rationale:

- SARIMAX models seasonality and external factors.
- Volume correlates with price.

2.9 Cell 9: Automated Model Selection with pmdarima

Purpose: Automate ARIMA/SARIMAX parameter selection.

Actions:

- Run auto_arima with s=12 and volume.
- Forecast with best model, compute RMSE, plot results.

Results:

- Best model: ARIMA(2,1,2)(1,0,0)[12] with intercept, AIC: 12449.551, BIC: 12486.550.
- Out-of-sample RMSE: 260.0435.
- Statistical tests: Ljung-Box (Q) = 0.03 (p = 0.86), Jarque-Bera (JB) = 3608.33 (p = 0.00), indicating non-normal residuals.

Rationale:

- auto_arima optimizes parameters.
- Stepwise search balances speed and accuracy.

2.10 Cell 10: SARIMA-GARCH Model for Volatility

Purpose: Combine SARIMA and GARCH for price and volatility forecasting.

Actions:

- Fit SARIMA(1,1,1)(1,1,0,12), extract residuals.
- Fit GARCH(1,1), forecast prices and volatility.
- Plot results with 95% confidence intervals.

Results:

- Out-of-sample RMSE: 442.6291.
- GARCH BIC: 12787.3.
- Volatility parameters: alpha[1] = 0.1508, beta[1] = 0.8175 (both significant).

Rationale:

- GARCH models heteroskedasticity.
- SARIMA captures price trends.

2.11 Cell 11: Model Evaluation and Comparison

Purpose: Compare models using metrics.

Actions:

- Compile RMSE, AIC, BIC.
- Create bar plots for RMSE and AIC/BIC.
- Plot all forecasts.

Results:

- Metrics table (see Cell 20 for updated values).
- Auto ARIMA/SARIMAX and LSTM outperformed simpler models.

Rationale:

- Comprehensive comparison identifies top performers.
- Visualizations aid interpretation.

2.12 Cell 12: Interactive Visualizations

Purpose: Create an interactive forecast dashboard.

Actions:

- Plot all model forecasts with unified hover mode.

Results:

- Dashboard consolidated results interactively.

Rationale:

- Hover mode enables precise comparisons.

2.13 Cell 13: Conclusion and Recommendations

Purpose: Summarize findings and propose next steps.

Actions:

- Summarize seasonality, performance, limitations.
- Recommend validation, exogenous variables, LSTM exploration.

Results:

- Seasonality: s=12 supported.
- Recommendations: Validate seasonality, add variables.

Rationale:

- Summary consolidates insights.
- Recommendations guide improvements.

2.14 Cell 14: Refine SARIMAX Model

Purpose: Optimize SARIMAX by testing seasonal orders.

Actions:

- Test orders: (1,1,0,12), (1,1,0,5), (0,0,0,0).
- Fit SARIMAX(1,1,1), select best by AIC.
- Plot best forecast, display results.

Results:

- Best model: SARIMAX(1,1,1)(0,0,0,0).
- AIC: 12434.5434, RMSE In: 57.6697, RMSE Out: 276.2554.
- Table:

Seasonal Order	AIC	BIC	RMSE In	RMSE Out
	12904.0664	12930.4075	66.5631	442.4877
(1,1,0,5)	14057.4702	14083.8599	82.6145	3.1358×10^{58}
(0,0,0,0)	12434.5434	12455.6800	57.6697	276.2554

- Note: The RMSE Out for (1,1,0,5) is anomalously high, likely due to numerical instability or inappropriate seasonality.

Rationale:

- Testing multiple orders validates seasonality.
- AIC balances fit and complexity.

2.15 Cell 15: Refine SARIMA-GARCH Model

Purpose: Tune GARCH parameters.

Actions:

- Fit SARIMA(1,1,1)(1,1,0,12), extract residuals.
- Grid search GARCH(p,q) for p,q=1,2.
- Select best by BIC, forecast, plot results.

Results:

- Best model: GARCH(1,1), BIC: 12787.3141, RMSE Out: 442.6291.
- Table:

GARCH Order	BIC	RMSE Out
	12787.3141	442.6291
(1,2)	12797.7082	442.6291
(2,1)	12788.8514	442.6291
(2,2)	12801.6157	442.6291

Rationale:

- Tuning optimizes volatility modeling.
- BIC ensures parsimony.

2.16 Cell 16: Fetch Real-Time Data

Purpose: Extend dataset with AAPL data.

Actions:

- Fetch AAPL data (2021-12-31 to 2025-05-08) via yfinance.
- Merge, clean, plot updated data.

Results:

- Data merged successfully, no duplicates.
- Updated data included additional columns (e.g., Daily Return, MA50).

Rationale:

- Real-time data ensures relevance.
- Cleaning maintains consistency.

2.17 Cell 17: Prepare Data for LSTM

Purpose: Create sequences for LSTM.

Actions:

- Split updated data (80% train, 20% test).
- Scale data, create 60-day sequences.
- Reshape for LSTM.

Results:

- Training sequences: (1400, 60, 1).
- Testing sequences: (305, 60, 1).

Rationale:

- Sequences capture temporal dependencies.
- 60-day length balances patterns and efficiency.

2.18 Cell 18: Train and Evaluate LSTM

Purpose: Implement and evaluate LSTM model.

Actions:

- Build LSTM: Two layers (50 units each), Dropout(0.2), Dense(25), Dense(1).
- Compile with Adam optimizer (learning rate = 0.001), loss: MSE.
- Train for 20 epochs, batch size 32, validation split 0.1.
- Compute RMSE, plot training history and predictions.

Results:

- Out-of-sample RMSE: 108.3767.
- Training showed decreasing loss (e.g., validation loss at Epoch 20: 0.00065).

Rationale:

- LSTM captures long-term dependencies.
- Dropout prevents overfitting.

2.19 Cell 19: Update Best Model

Purpose: Retrain best model on updated data.

Actions:

- Select best model: LSTM (RMSE: 108.3767).
- Retrain on updated data, compute RMSE, plot forecast.

Results:

- Updated LSTM RMSE: 96.1212.

Rationale:

- Updating ensures market relevance.
- Consistent evaluation maintains comparability.

2.20 Cell 20: Compare All Models

Purpose: Final model comparison.

Actions:

- Update metrics table.
- Create bar plots for RMSE, AIC/BIC.
- Plot all forecasts.

Results:

- Metrics table:

Model	RMSE In	RMSE Out	AIC	BIC
AR(2)	17.2755	264.6477	-8438.0591	-8416.9199
ARIMA(1,1,1)	17.2872	278.4383	12462.6640	12478.5205
Auto ARIMA/SARIMAX	NaN	260.0435	12449.5510	12486.5500
SARIMAX Refined	57.6697	276.2554	12434.5434	12455.6800
SARIMA-GARCH	NaN	442.6291	NaN	12787.3141
LSTM	NaN	108.3767	NaN	NaN
Updated LSTM	NaN	96.1212	NaN	NaN

- Note: AR(2) AIC is unusually negative, possibly due to scaling; requires validation.

Rationale:

- Comprehensive comparison highlights LSTM's superiority.

2.21 Cell 21: Final Visualizations

Purpose: Create final interactive dashboard.

Actions:

- Plot all forecasts with unified hover mode.

Results:

- Dashboard showed LSTM's close alignment with actuals.

Rationale:

- Interactive visualization aids stakeholder engagement.

2.22 Cell 22: Final Conclusion

Purpose: Summarize project and outline future work.

Actions:

- Highlight achievements: refined models, LSTM, real-time data.
- Propose hybrid models, more variables, automation.
- Note limitations: LSTM tuning, data dependencies.

Results:

- Achievements: Auto ARIMA/SARIMAX RMSE = 260.0435, SARIMAX Refined RMSE = 276.2554, SARIMA-GARCH RMSE = 442.6291, Updated LSTM RMSE = 96.1212.
- Future work: Hybrid models, sentiment analysis, automation.

Rationale:

- Summary emphasizes success.
- Future work guides enhancements.

3 Key Findings

- Seasonality: Monthly (s=12) supported by ACF/PACF, though refined SARIMAX favored no seasonality (0,0,0,0).
- Performance: Updated LSTM (RMSE: 96.1212) and Auto ARIMA/SARIMAX (RMSE: 260.0435) outperformed others.
- Volatility: SARIMA-GARCH improved confidence intervals but had higher RMSE (442.6291).
- Real-Time: yfinance integrated AAPL data effectively.
- Visualization: Plotly dashboards enhanced interpretability.

4 Recommendations

- Deploy Updated LSTM or Auto ARIMA/SARIMAX with automated retraining.

- Investigate SARIMAX(1,1,1)(1,1,0,5) instability (RMSE: 3.1358×10^{58}).
- Add exogenous variables (e.g., market indices, X post sentiment).
- Explore SARIMAX-LSTM hybrids and Transformer models (e.g., TimeGPT).
- Optimize computational efficiency for scalability.

5 Conclusion

The stock price forecasting project successfully developed a suite of models, with the Updated LSTM achieving the best performance (RMSE: 96.1212). Real-time data integration and interactive visualizations enhanced applicability. Future work should focus on hybrid models, addressing numerical instabilities, and scalable pipelines.