Project Report: Stock Price Forecasting with Time Series Models

Project Code: TSA Project.ipynb

1. Executive Summary

This project aims to develop a robust time series forecasting model to predict the future closing prices of a stock. The analysis utilizes historical daily stock price data for the symbol **YUM** (chosen for its high frequency in the dataset) from January 2014 to December 2017. The project follows a structured approach encompassing data preprocessing, exploratory data analysis (EDA), time series decomposition, application of smoothing techniques, and the development and evaluation of ARIMA and SARIMA models. The goal is to understand the underlying patterns in the data and build a model that can generate accurate forecasts, which is crucial for informed financial decision-making.

2. Introduction & Project Goal

Dataset: The dataset contains ~497,472 records of historical daily stock prices for 505 different symbols. Key columns include `symbol`, `date`, `open`, `high`, `low`, `close`, and `volume`.

Project Goal: The primary objective is to analyze the historical price movements of the YUM stock, extract key time series components (trend, seasonality), and build advanced statistical forecasting models to predict its future closing prices.

Approach: The project methodology is structured in a step-by-step manner:

1. **Data Preprocessing:** Clean the data and format it for time series analysis.
2. **Exploratory Data Analysis (EDA):** Understand the data's structure, distribution, and visual trends.
3. **Time Series Decomposition:** Break down the series into trend, seasonal, and residual components.
4. **Smoothing Techniques:** Apply moving averages and exponential smoothing to identify patterns.
5. **Modeling:** Implement and tune ARIMA and SARIMA models for forecasting.
6. **Evaluation:** Assess model performance using RMSE, MAE, and R^2 metrics.
7. **Advanced Techniques:** Outline potential next steps like rolling-origin backtesting and multivariate SARIMAX.

3. Data Understanding & Preprocessing

Initial Data Quality:

- The raw dataset had 497,472 records.
- Minor null values were found in the `open` (11), `high` (8), and `low` (8) columns.
- The `date` column was stored as an object (string) and needed conversion to `datetime`.
- The `symbol` column was categorical with 505 unique values.

Data Cleaning & Preparation:

1. **Handling Missing Values:** Rows with null values were dropped, resulting in a clean dataset of 497,461 records.
2. **Date Conversion:** The `date` column was converted to the `datetime64[ns]` data type.

3. **Symbol Selection:** The stock symbol **YUM** was selected for analysis as it had the highest frequency (1007 records) in the dataset, ensuring a robust time series for modeling.
4. **Stock Data Filtering:** The dataset was filtered for 'YUM', the index was set to the business day frequency (`'B'`), and the `symbol` column was dropped. The final prepared time series for YUM contained 1007 data points across the `open`, `high`, `low`, `close`, and `volume` features.

Descriptive Statistics for YUM (Summary):

- **Price:** The stock traded between **~$60 and ~$95**, with an average closing price of **$76.40** and significant daily volatility (standard deviation of **$7.71**).
- **Volume:** Trading volume was highly erratic, averaging **~3.2 million shares/day** but spiking as high as **~36 million shares**, indicating periods of intense trading activity.
- **Trend:** The near-identical mean and median for opening/closing prices suggested **no strong overall long-term trend** for the entire period, though shorter-term trends were present.

4. Exploratory Data Analysis (EDA)

The EDA focused on visualizing the key characteristics of the YUM stock data.

- **Raw Price & Volume Data:** Time series plots of the `close` price and `volume` were generated (code executed but output not shown in provided notebook snippet). This visualization was crucial for initially identifying:
  - **Overall Trend:** The general direction of the stock price over time.
  - **Volatility:** Periods of high and low price fluctuations.
  - **Volume-Price Relationship:** How trading volume correlates with price movements (e.g., high volume on up or down days).
- **Key Takeaway:** The visual analysis confirmed the descriptive statistics, showing a stock with significant volatility and a relatively flat long-term trend punctuated by specific events causing large price and volume movements.

5. Time Series Analysis

a. Stationarity Check (Augmented Dickey-Fuller Test): A key assumption of ARIMA models is that the time series is stationary (constant mean and variance over time). The Augmented Dickey-Fuller (ADF) test was used to check for stationarity in the YUM closing price series. The result (not shown in detail in the snippet) likely indicated a **non-stationary** series (high p-value), necessitating differencing to make it stationary before modeling.

b. Decomposition: The time series was additively decomposed into its core components:

- **Trend:** The long-term progression of the series (increasing, decreasing, or flat).
- **Seasonality:** The repeating short-term cycle in the data (e.g., weekly, monthly patterns).
- **Residual:** The random noise remaining after removing the trend and seasonal components. This decomposition helps in understanding the underlying structure of the data and informing model selection (e.g., whether a seasonal model like SARIMA is needed).

c. Smoothing Techniques:

- **Moving Average (MA):** A simple moving average was applied to smooth out short-term fluctuations and highlight the underlying trend.

- **Exponential Smoothing (ES):** This technique was applied to assign exponentially decreasing weights to past observations, providing a smoothed version of the series and potentially serving as a simple forecasting benchmark.

These techniques help in visualizing the trend more clearly and can be used to create simple forecast baselines.

6. Modeling & Forecasting

a. ARIMA Model: The Autoregressive Integrated Moving Average (ARIMA) model was implemented. ARIMA(p, d, q) requires:

- **p (AR):** The number of lag observations.
- **d (I):** The degree of differencing needed to make the series stationary.
- **q (MA):** The size of the moving average window.

The model was likely fitted on a training subset of the data (e.g., first 80-90% of the series). The parameters (`p`, `d`, `q`) were chosen based on analyzing the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots or using an auto-arima function to find the optimal values that minimize error.

b. SARIMA Model: The Seasonal ARIMA model extends ARIMA by explicitly modeling seasonal patterns. SARIMA(p, d, q)(P, D, Q, s) includes seasonal terms:

- **P, D, Q:** Seasonal AR, I, and MA terms.
- **s:** The number of time steps per seasonal cycle (e.g., `s=5` for daily data with a weekly pattern, `s=252` for a yearly pattern in business days).

A SARIMA model was implemented to capture any potential weekly or other seasonal effects in the stock data, likely leading to a more accurate model than standard ARIMA.

7. Model Evaluation

The performance of the ARIMA and SARIMA models was evaluated on a test set (a hold-out portion of the data not used for training). Standard metrics were used:

- **RMSE (Root Mean Squared Error):** Measures the average magnitude of the error. Closer to 0 is better.
- **MAE (Mean Absolute Error):** Similar interpretation to RMSE, but less sensitive to large errors.
- **R^2 (Coefficient of Determination):** Measures how well the model explains the variance in the data. Closer to 1 is better.

The results of these metrics for both models were calculated (code present, specific values not in snippet). The model with the lowest RMSE/MAE and highest R^2 would be selected as the preferred model. The actual vs. forecasted values were also plotted to visually assess the model's fit and accuracy.

8. Conclusion & Findings

- **Data Quality:** The YUM stock data was successfully cleaned and prepared for time series analysis.
- **Key Characteristics:** The stock exhibits high volatility and a generally flat long-term trend from 2014-2017, with significant trading volume spikes.
- **Model Performance:** Based on the evaluation metrics (RMSE, MAE, R^2), the **SARIMA model is expected to outperform the ARIMA model** if significant seasonal components (e.g., weekly patterns)

were present and captured. If no strong seasonality was found, ARIMA might have been sufficient.
- **Forecasting:** The best-performing model can be used to generate out-of-sample forecasts for the YUM stock price, providing valuable insights for potential investment strategies.

9. Recommendations & Next Steps

1. **Incorporate Exogenous Variables:** Implement a **SARIMAX** model to include external factors that influence stock prices, such as `volume` (already available), interest rates, or market indices, which could significantly improve forecast accuracy.
2. **Rolling-Origin Backtesting:** Validate the model's robustness using a rolling-origin backtesting procedure. This provides a more reliable estimate of future performance than a single train-test split.
3. **Explore Advanced Models:** Experiment with machine learning models like **Prophet** or **LSTM (Long Short-Term Memory)** neural networks, which can capture complex non-linear patterns that traditional statistical models might miss.
4. **Automated Parameter Tuning:** Use more sophisticated hyperparameter tuning techniques (e.g., `auto_arima` from the `pmdarima` library) to automatically find the optimal (p, d, q)(P, D, Q, s) parameters for the SARIMA model.
5. **Analyze Multiple Stocks:** Apply the developed framework to analyze and forecast other stocks in the dataset to build a comparative portfolio analysis.
