**Bull Market or Bear Market: Time Series Price Prediction for Q1 2024:**

**Data Understanding and Data Preparation**

John Vincent Deniega, Ravita Kartawinata, Gabi Rivera

Master of Science in Applied Data Science, University of San Diego

**Data Understanding**

***Collect Initial Data***

Stock market data was retrieved via Aroussi’s (2023) “yfinance” Python application programming interface (API) that sources Yahoo Finance information updated during normal trading hours. The API ticker function retrieves comprehensive daily logs of the stock market movements and metadata from opening prices, high, low, and closing prices as well as volume. Opening, closing, and volume information seem to be the most promising on initial review of the vectors that may be called. Additional methods regarding including data calls on news, holders, and other information is beyond the scope of this initial price prediction objective. Data appears to be complete with a small operational test for API functionality and any missing values that may arise may be handled during subsequent data preparation and modeling phases.

***Describe Data***

Data called by the *history* method retrieves required stock price data in a Pandas DataFrame format. Columns of interest include price information that is returned as a Series containing stock prices as floating-point type numbers. Resulting DataFrame returns an index of *pandas.Timestamp* type objects that may require further processing for more compatible time-series analysis. Due to the API nature of a live data source, the size of the total dataset changes daily, though the DataFrames and Series that may be instantiated will change depending on the time of API call unless a time period is explicitly specified.

***Explore Data***

Initial DataFrames were created off of the S&P 500 exchange-traded fund (ETF) with ticker symbol, “SPY” and Amazon.com Incorporated (AMZN). Determination of stationarity status via the Augmented Dickey-Fuller (ADF) test was employed with .05 significance level initially set for analysis. ADF test was applied to the raw time series data of SPY and AMZN, which determined that these series do not exhibit stationarity, and so subsequent time series transformation is required for modeling as depicted in Figures 1a and 2a, the original time series. Figures 1b and 2b depict the plot after transformation using subsequent Seasonal-Trend Decomposition using Locally Estimated Scatterplot Smoothing (LOESS) (STL) regression. SPY and AMZN’s trend and seasonal components were clearly parsed in both cases. First-degree differencing was used to remove both trend and seasonal components for further model exploration of the remaining residual time series. ADF tests of p-values lower than .05 significance level confirm that the time series that were transformed in this way resulted in time-series stationarity. Figures 1c-d and 2c-d demonstrate autocorrelation (ACF) and partial correlation (PACF) plots of SPY and AMZN. Target time series’, SPY’s, ACF suggests moving average (MA) orders of 0, 1, 2, 11, 13, 14, 21, 37, and 40 lag periods are candidates for ARIMA model development. PACF suggests autocorrelation (AR) orders of 1, 2, 3, 8, 9, 11, 14, 21, 58, 63, and 70 lag periods are candidates for ARIMA model development.

**Figure 1**

*SPY Exploratory Data Analysis Plots*

A graph showing a line graph

Description automatically generatedA graph of a graph

Description automatically generated with medium confidence A graph with blue dots

Description automatically generated A graph with blue dots

Description automatically generated

*Note.* (A. Top left) Time Series Plot of SPY 5-year Closing Price. (B. Top Right) STL decomposision of SPY time series using LOESS technique. (C. Bottom Left) SPY ACF plot. (D. Bottom Right) SPY PACF plot.

**Figure 2**

*AMZN Exploratory Data Analysis Plots*

A graph of a stock market

Description automatically generatedA line graph of different types of data

Description automatically generated with medium confidence

A graph with blue dots

Description automatically generated A graph with blue dots and numbers

Description automatically generated

*Note.* (A. Top left) Time Series Plot of AMZN 5-year Closing Price. (B. Top Right) AMZN decomposision of AMZN time series using LOESS technique. (C. Bottom Left) AMZN ACF plot. (D. Bottom Right) AMZN PACF plot.

***Verify Data Quality***

No missing data values were observed during initial time series plotting approach on market days. Missing values in terms of the daily nature of the time series occurred on weekends and holiday dates where stock trading does not typically occur. These missing values were then imputed and filled in forward by propagating the last valid value to the next. Then, the time series datasets were explicitly labeled as a daily frequency time series for model development and fit purposes.

**Data Preparation**

***Select Data***

SPY and AMZN are the time series of interest for the purposes of this project. Price-related information such as opening price, closing price, volume information as obtained by the *history* method call and resulting DataFrame are the primary data sources to form time series for model development.

***Clean Data***

As previously mentioned, because the stock market and other trading markets typically do not trade daily and typically trade on a normal Monday to Friday weekly schedule, missing time series data on weekends and holidays will be imputed using the last known valid value.

***Construct Data***

Additional columns are expected to be derived from original columns for model development. A column containing the difference between the closing price and the opening price of a security during that period will be derived from the existing aforementioned columns. From this derived column, an additional column containing two-levels to represent the directionality of a time series relative to its previous period will be used to represent whether a security increases or decreases in price for that period.

***Integrate Data***

It is expected that individual stocks for comparison, such as AMZN, may be considered as external predictors that may be included in a SPY model in order to determine its efficacy in predictive power. As API populates additional data from Yahoo Finance, the time series is expected to be dynamic and the subsequent models auto-regressing or shifting observation windows with the most current additional information concatenated to the end of the window where beginning is shifted forward, as appropriate

***Format Data***

Time series data to be used for model instantiation and fit will be locally stored in a DataFrame or Series, as appropriate. Models will be coded explicitly as daily time series in sequential order starting from earliest period to most recent period. Training periods shall occur prior to validation periods and both periods shall occur prior to any future or test periods, as appropriate.

**References**

Aroussi, R. (2023). *yfinance* (Version 0.2.32). PyPI. **https://pypi.org/project/yfinance/**