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

**Modeling and Evaluation**

John Vincent Deniega, Ravita Kartawinata, Gabi Rivera

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

**Modeling**

***Selecting Modeling Techniques***

Daily stock market logs were mined at the time of API call. Data and associated data types that were retrieved consisted of stock prices (numerical float), volume (integer), and time (Pandas datetime). An additional variable of interest was derived from stock prices to indicate whether or not the daily price for the time period queried resulted in a positive net increase in price (binary). This derived variable serves as the dependent *‘y’* variable for a logistic regression model to be trained to provide a forecast targeting price increase logic.

In order to accomplish this, additional variable predictors were derived as intermediates in the process of fitting the logistic regression model to accomplish this binary outcome. These new predictors are *‘open\_close’* as well as *‘high\_low’* which are used to represent the difference between the two original predictors as a combined item in both cases. The *‘open\_close’* stock price was selected as the most complete predictor of the period that smoothed out intraday price fluctuations. This thereby simplified the numerical difference into a binary outcome field, *‘positive’*, as either *‘0’* or *‘1’* with the latter indicating that the price increased during that discrete period. The binary transformation is used intently to capture the closing stock market price behavior as positive signifying up and negative to down. The time series data – now with additionally constructed features – were then differenced in order to introduce stationarity into the resulting time series for modeling. DataFrames at lag periods one through five were separately constructed for their respective logistic regression model in order to select the best performing lag period and other parameters against the validation set.

In comparison to other methods, the original time series data types pulled from the *yfinance* API were subjected to first-ordered differencing of the closing price predictor before feeding the data into Autoregressive Integrated Moving Average (ARIMA), Simple Exponential Smoothing (SES), and Advanced Exponential Smoothing (AES) methods. The assumption is that the time series exhibits stationarity to remove the influences of both trend and seasonality from the methods. The goal of these aforementioned methods is to predict the closing stock prices at a given target future date solely based off of historical data. Inspection of the data’s autocorrelation plot in Figure 1 revealed potentially significant moving average orders at periods 0, 1, 2, 11, 13, 14, 21, 37, and 40. Subsequent inspection of the data’s partial autocorrelation plot in Figure 2 revealed potentially significant autoregression orders at periods 1, 2, 3, 8, 9, 11, 14, 21, 58, 63, and 70. This information was key to determining which parameters to iterate through in order to discover the highest performing parameters for lag and moving average orders of ARIMA and AES’s individual models.

***Generating a Test Design***

As mentioned earlier, both model-based and data-driven methods, ARIMA, AES, & SES, were used to develop price predictors and logistic regression models to represent stock market movement through price prediction and binary classification. For the price prediction objective, statistical measurements such as root mean square error (RMSE) and mean absolute percentage error (MAPE) were used to evaluate predictive performance between model predictions and truly observed values due to their interpretable values in being in the same unit as the original series as well as a relative proportional measurement as a percentage of the error, respectively.

Additionally, in order to evaluate with consideration for both fitness of the data with complexity of the method used, Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) were considered when selecting the best method to deploy on unseen data. Of these two criteria, a lower number of parameters and a higher log likelihood of fitting the data aim to mitigate the effects of overfitting a model to the training data.

For binary forecasting, a confusion matrix provided performance metrics for the purpose of evaluating the predictive performance of the classifier against the validation data. Precision, sensitivity, F1 score, and accuracy were included in the confusion matrix to assist in different models matching different investment risk profiles. For more conservative risk profiles, high precision models may be preferred due to their focus on higher probability positive price movements. For more risk-tolerant and higher return risk profiles, high accuracy models may be preferred so as to optimize for higher recall on positive signals, but counter-balanced with the higher specificity on non-positive signals (assuming equally weighted costs on both positive and non-positive periods).

***Building the Models***

SES, AES, ARIMA, and logistic regression forecasting methods were explored throughout the model building process. Among the price predictors, the data-driven AES method exhibited the highest performance based on RMSE and MAPE. Among binary forecasters, logistic regression exhibited the highest performance based on confusion matrix results. With respect to their respective parameters, AES was set to have multiplicative damped trend, additive seasonality, three periods per season, and a heuristic initiation method. Fit was set at 0.1 for both smoothing level and trend parameters. These were determined by running AES through multiple combinations of each parameter such that for both trend and seasonal parameters ‘*add, mul, additive, multiplicative, and none*’ were iterated through to find the combination that yielded the best performing metrics. The same process was performed for parameters in damped trend, seasonal period length, initiation methods, and multiple fits. Then, each parameter was subjected to predictive performance criteria sequentially. The validation dataset used for the calculation of the performance criteria is set from the last year period and ending on November 27, 2023. The results are further discussed under assessment and evaluation.

As for logistic regression, the *‘positive’* field was set as the dependent outcome variable. The stock price predictors along with the new predictors derived by lagged periods were differenced at three periods before feeding into the logistic regression model. Then, validation data was set to the next 200 periods after the training model, resulting in the confusion matrix found in Figure 6.

***Model Assessment and Evaluation***

Figure 3 shows SES model forecast with poor predictive curve spanning across the trailing year range. The AIC score was recorded at 3042 while BIC was recorded at 3051, which is to be compared with models throughout this discussion. RMSE was recorded at 31.71 and MAPE was recorded at .06. This may be interpreted as an error of $31.71 USD and a 6% price deviation.

Figure 4 shows ARIMA model forecast performing poorly in predicting the validation set similar with SES plot. The forecast curve flattened out throughout the last year prediction range. For performance metrics review, optimal ‘*pdq’* parameters of 14, 1, 1, respectively, yielded an AIC of 10377 and a BIC of 10466. The RMSE was recorded at 418.23 and the MAPE was recorded at 1.0. This may be interpreted as an error of $418.23 USD and a 100% price deviation, suggesting that even with parameter optimization, ARIMA may be unfeasible for price prediction purposes.

As shown in Figure 5, the AES method was able to forecast the exponential pattern compared to the validation data. Although, the trend and seasonality patterns were not explicitly reflected to also match the actual validation results it appears that the smoothing sufficiently compensated for increased predictive performance. Among all of the price prediction models, the data-driven AES method performed the best against the validation set with half of the error of SES for only a one-fifth higher information criterion with AIC at 3884 and BIC at 3923. The RMSE was recorded at 15.06 and the MAPE was recorded at .03. This may be interpreted as an impressive minimal error of $15.06 USD and a 3% price deviation as of November 27, 2023’s data retrieval depicted in Figure 7.

The final model is of a logistic regression with resulting confusion matrix shown in Figure 6. The overall accuracy is at .62 with precision at .77 and sensitivity at .61. This may be interpreted as a model that is likely best suited for more conservative risk profiled investors due to its higher precision performance.

**Figure 1**

*Autocorrelation Plot on SPY (S&P 500 ETF)*

*A graph with blue dots

Description automatically generated*

**Figure 2**

*Partial Autocorrelation Plot on SPY (S&P 500 ETF)*

**A graph with blue dots

Description automatically generated**

**Figure 3**

*SES Forecast Model Plot*

A graph with blue lines

Description automatically generated

**Figure 4**

*ARIMA Forecast Model Plot*

A graph with orange and blue lines

Description automatically generated

**Figure 5**

*AES Forecast Model Plot*

*A graph with orange and blue lines

Description automatically generated*

**Figure 6**

*Logistic Regression Confusion Matrix*

A screenshot of a graph

Description automatically generated

**Figure 7**

*Final Advanced Exponential Smoothing Model with Optimal Parameter Values*

**A screenshot of a computer program

Description automatically generated**

**References**

Industrial Business Machines Corporation (2021). *Introduction to CRISP-DM*. Industrial Business Machines Corporation. **https://www.ibm.com/docs/en/spss-modeler/saas**

**?topic=guide-introduction-crisp-dm**