**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***

The data types used to mine the stock market daily logs are set as numerical float for stock prices, integer for volume, datetime for the timestamps, and a binary data type was created to satisfy the logistic regression model. Additional predictors were created in the process of building the logistic regression model to define the classification forecasting 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 closing stock price was selected as the most representative predictor and so the binary outcome field was created by setting the open\_close field as either 0 or 1 at 0.5 threshold. The binary transformation is used intently to capture the closing stock market price behavior as positive signifying up and negative to down. Then as a last modeling technique for logistic regression, the predictors were differenced and lags 1-3 were taken to select the best performing parameters.

In comparison to other methods, the original time series data types pulled from YF library were subjected to first ordered differencing of the closing price predictor before feeding the data to ARIMA and Exponential Smoothing (AES & SES) methods. The assumption is that the timeseries is at a stationary course to remove the influences of trend and seasonality from the forecasting methods. The goal of these mentioned methods is to predict the closing stock prices in a short time period based off of historical residual data. To run the best price prediction models then, autocorrelation and partial autocorrelation were the key players in determining the highest performing parameters for lag and moving average orders of ARIMA and AES’s seasonal periods.

***Generating a Test Design***

As mentioned earlier, there were two groups of modeling methods used which are ARIMA, AES, & SES to represent price predictors and logistic regression to represent the stock market fluctuation through classification. For the price predictors, the criteria’s to test for goodness of the model and the parameters that was elected are decidedly through mean square error (MSE), mean absolute percentage error (MAPE), Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC) scoring. MSE is used to check the accuracy of the forecasted values from the actual. The lower the MSE value is, then the closer the prediction is to the actual value. MAPE will be used to measure the effectiveness of the price predictor models. It measures the averaged percent error of each entry in the system. AIC and BIC are favored at their lowest to help determine the models that have the lowest number of parameters with the highest log likelihood for an ideally fitted model.

For logistic regression, a confusion matrix was utilized to calculate the performance metrics of the classification method and the parameters applied. Precision, sensitivity, and accuracy scores are the main criteria for the model’s success. Precision provides the ratio of true positive from the total predicted positive. Sensitivity provides the ratio of true positive form the total actual positive. Both of these metrics alongside accuracy are used to conceptualize the highest performing model parameters and ultimately, the outcome of the forecast.

***Building the Models***

SES, AES, ARIMA, and logistic regression forecasting methods were explored throughout the model building process. Among the price predictors, AES method performed the best as well as logistic regression for classification. These two models are high lightened in this section instead. In terms of parameters, AES was set to have multiplicative trend, damped trend, additive seasonal, 3 seasonal periods, and heuristic initiation method. Fit was set to have 0.1 smoothing level and trend. These were determined by running AES through multiple combination of each parameter such that for trend/seasonal ‘*add, mul, additive, multiplicative, and none*’ were looped against each other as an example. The same process was performed for damped trend, seasonal periods, initiation methods, and fits. Then, each parameter was subjected to predictive performance criteria set under test design to select the best combinations. The validation dataset used for the calculation of the performance criteria is set from the last year period and ending on 27Nov2023. The results are discussed under assessment and evaluation.

As for logistic regression, the model was built straight forward. The created positive field was set as the outcome variable. The stock price predictors along with the new predictors were difference at lag 3 before feeding into the logistic model. Then validation data was set at 200 period after the training model to produce the confusion matrix in Figure 4.

***Model Assessment and Evaluation***

Figure 1 shows SES model forecast with poor predictive curve spanning across the last year range. The AIC score is at 3042 while BIC is fairly similar at 3051 score. MSE is at 1005 with MAPE score at 6%.

Figure 2 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 metrics performance review, optimal pdq parameters of 14, 1, 1 yielded an AIC of 10377 and a BIC of 10466. The MSE is at 174914 and a MAPE of 100%

As shown in Figure 3, the AES method was able to forecast the exponential pattern compared to the validation data. Although, the trend and seasonality patterns were not reflected to also match the actual validation results. Amongst all of the price prediction model however, AES visually performed the best. The evaluation metrics had the lowest AIC at 3566 and BIC at 3600 indicating a better fitted model compared to other AES model parameter iteration. The MSE was also 348 and a MAPE of 3%.

The last model is logistic regression and shown in Figure 4 is the result of the confusion matrix with predictors set at lag-3. The overall accuracy is at 62% with precision at 77% and sensitivity at 61%.

**Figure 1**

*SES Forecast Model Plot*

A graph with blue lines

Description automatically generated

**Figure 2**

*ARIMA Forecast Model Plot*

A graph with orange and blue lines

Description automatically generated

**Figure 3**

*AES Forecast Model Plot*

*A graph with orange and blue lines

Description automatically generated*

**Figure 4**

*Logistic regression confusion matrix*

A screenshot of a graph

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**