0.0.1 Performance Metrics for Evaluation

To assess the quality and accuracy of forecasting models, it is essential to employ performance metrics that reflect how closely predicted values align with actual observations. There is no single metric universally accepted as the best for evaluating forecasting accuracy. Nevertheless, accuracy is widely regarded as the primary criterion for assessing the quality of a forecast, as it reflects how closely the predicted values match the actual ones. The goal of forecasting is to minimize the forecast error, defined as the difference between actual and predicted values [Kisang and Alfonso, 2003].
In the following section, the key evaluation metrics employed in this study are presented and briefly discussed.

The Root Mean Squared Error (RMSE):
Root Mean Squared Error (RMSE) is frequently employed to assess the predictive performance of a model by evaluating the typical size of its errors. It is derived by taking the square root of the average squared difference between the estimated and actual values. As it retains the same units as the response variable, RMSE offers a meaningful indication of the model's overall accuracy [Gifty and Li, 2024]:

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(Y_t - \hat{Y}_t)^2}$$

where:

- $\hat{Y}_t$ = the predicted value for time period t
- Y_t = the actual observed value in time period t
- n = the number of forecast observations in the evaluation period [Kisang and Alfonso, 2003].

Mean Absolute Error (MAE):
The Mean Absolute Error (MAE) is a commonly used metric for evaluating the average size of prediction errors. It is computed as the mean of the absolute differences between predicted and actual values, providing a measure of error magnitude without considering the direction of the deviations [Gifty and Li, 2024]:

$$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n}\left|Y_t - \hat{Y}_t\right| \quad \text{[Kisang and Alfonso, 2003]}$$

When prediction errors follow a normal distribution, RMSE is the most appropriate metric, as it corresponds to the maximum likelihood estimator under Gaussian assumptions. If errors follow a Laplacian distribution (a symmetric heavy-tailed alternative to the normal), MAE becomes more appropriate. These choices align with the statistical properties of the data. However, in many real-world applications, errors are neither perfectly normal nor independent, so RMSE and MAE may not always be ideal [Hodson, 2022]. To make the evaluation more robust to such violations, an additional metric—Median Absolute

Deviation (MAD)—can be used [Hodson, 2022].

Median Absolute Deviation (MAD):
MAD is less sensitive to outliers and non-normality, since it measures the deviation from the median (rather than the mean, which is more influenced by outliers), making it valuable when the distribution of errors deviates from classical assumptions such as normality [Hodson, 2022]. This metric is computed by taking the average of the absolute differences between the actual and predicted values across all time periods. The formula is:

$$MAD = \frac{1}{n} \sum_{t=1}^{n} |Y_t - \hat{Y}_t|$$

If the forecast perfectly matches the actual data, the MAD will be zero. A larger MAD indicates greater forecast error. Thus, when comparing different forecasting methods, the one with the lowest MAD is generally considered the most accurate [Kisang and Alfonso, 2003].

Since we aim to evaluate model accuracy, both the Mean Absolute Error (MAE) and the Median Absolute Deviation (MAD) provide complementary insights. MAE measures the average magnitude of prediction errors, making it a widely used indicator of overall accuracy. MAD, as the median of absolute errors, captures the typical prediction error and is less sensitive to extreme values. When both MAE and MAD are low and approximately equal (MAE $\approx$ MAD), it suggests that the model consistently produces small and accurate predictions. Conversely, if MAE is substantially greater than MAD, it indicates that the model occasionally generates large errors. Therefore, analyzing both metrics jointly offers a more robust evaluation of forecasting accuracy by balancing average performance with error consistency.

Direction Accuracy (DA):
Since the objective is to forecast asset returns, it is crucial to assess how frequently the model correctly predicts the direction of the return movement.
Directional Accuracy measures how often a model correctly predicts the direction (sign) of the change — not the magnitude — in a time series. It is particularly useful in financial forecasting, where predicting whether the market will go up or down is often more important than predicting the exact value.

$$\text{Directional Accuracy} = \frac{1}{n} \sum_{t=1}^{n} \mathbb{I}\left[\text{sign}(Y_t - Y_{t-1}) = \text{sign}(\hat{Y}_t - Y_{t-1})\right]$$

- Y_t: actual value at time t

- $\hat{Y}_t$: predicted value at time t

- $\mathbb{I}[\cdot]$: indicator function (1 if true, 0 if false)

- n: total number of predictions

The formula for Directional Accuracy used in this work was found on Wikipedia[1]. Although Wikipedia is generally not considered a reliable source for scientific research—due to the collaborative and non-peer-reviewed nature of its content—the formula itself is straightforward and intuitively understandable. Given its clarity and simplicity, and the fact that it is widely accepted in practice, I deemed it appropriate to use this expression without consulting additional academic references.

Coefficient of Determination (R^2):
The Coefficient of Determination (R^2) measures how well the independent variables explain the variation in the dependent variable. It ranges from 0 to 1, with 1 indicating a perfect fit and 0 suggesting no explanatory power [Gifty and Li, 2024].

$$R^2 = 1 - \left(\frac{\text{SSR}}{\text{SST}}\right)$$

SSR (Sum of Squared Residuals) is the sum of the squared differences between the predicted values and the mean of the dependent variable. SST (Total Sum of Squares) is the sum of the squared differences between the actual values and the mean of the dependent variable [Gifty and Li, 2024].

To conclude the section on performance metrics, and in order to identify the most accurate models, I adopt a ranking-based evaluation approach in line with [Kisang and Alfonso, 2003]. Specifically, a model ranking system is implemented to determine those models that exhibit superior predictive performance. For metrics where lower values correspond to better accuracy—such as RMSE and MAE—models are ranked in ascending order.

[1] https://en.wikipedia.org/wiki/Mean_directional_accuracy?utm_source=chatgpt.com

Bibliography

Aiyegbeni Gifty and Dr. Yang Li. A comparative analysis of lstm, arima, xgboost algorithms in predicting stock price direction. *Engineering and Technology Journal*, 09(08): 4978–4986, 2024. ISSN 2456-3358. doi: 10.47191/etj/v9i08.50. I.F. – 8.227.

Timothy O. Hodson. Root-mean-square error (rmse) or mean absolute error (mae): when to use them or not. *Geoscientific Model Development*, 15:5481–5487, 2022. doi: 10.5194/gmd-15-5481-2022. URL `https://doi.org/10.5194/gmd-15-5481-2022`.

Ryu Kisang and Sanchez Alfonso. The evaluation of forecasting methods at an institutional foodservice dining facility. *The Journal of Hospitality Financial Management*, 11(1):27–45, 2003. doi: 10.1080/10913211.2003.10653769. URL `https://doi.org/10.1080/10913211.2003.10653769`.