POLITECNICO MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE E DELL'INFORMAZIONE

Project report

Econometrics Project: Forecasting the NIFTY 50 Index to Develop a Trading Algorithm

Econometrics

Group 01

Clément Abissi (275788)
Théo Boltz (275796)
Mickael Teboul (285210)
Ulysse Grimault (274217)

Academic year: 2024/2025

Contents

1. Introduction And Literature Review

In recent years, algorithmic trading has become an essential component of financial markets, with increasing reliance on data-driven models to make investment decisions. This project aims to explore the predictive capabilities of econometric techniques to forecast the NIFTY 50 index, a major benchmark of the Indian equity market. By leveraging both univariate and multivariate time series models, we seek to develop a simple yet effective trading algorithm that can outperform a passive investment strategy. Based on existing literature and financial theory, we test our forecasts and trading rules under realistic market assumptions, including the incorporation of the Nikkei 225 index and the VIX volatility index as exogenous variables. Our ultimate goal is to demonstrate that statistical forecasting, when combined with financial reasoning, can provide a competitive edge in the market.

2. Data and Methodology

2.1. Dataset

To build our forecasting models and trading algorithms, we collect historical financial data from publicly available sources. Our main data source for index time series is the yfinance Python library, which provides access to Yahoo Finance's API. Using yfinance, we download daily closing prices for the NIFTY 50 index, representing the Indian stock market, and for the Nikkei 225 index, Japan's main equity benchmark. We select this indices due to their economic relevance and complementary trading hours, which allow us to investigate potential lead-lag relationships.

In addition to these indices, we includ the India VIX Index as a proxy for global market volatility. The VIX data is retrieved from an online source (investing.com), as it is not available via yfinance (only the one from yesterday is available). Including the VIX enables us to capture periods of heightened market uncertainty, which may influence the predictive power of our models and the behavior of the trading algorithm. Since the VIX open is the only data we are able to retrieve online (the close is not available), we will assume that the previous day's close corresponds to the current day's open for the purposes of our forecast.

The dataset covers a period of approximatley seven years, from June 2018 to June 2025, encompassing both stable and turbulent market conditions, including the COVID-19 crisis. All time series are aligned on a common calendar and clean to remove missing values or inconsistent data entries. This wide time span ensures robustness in both model training and out-of-sample evaluation.

We choose to work with daily frequency data instead of high-frequency (e.g., second-by-second) data for both practical and theoretical reasons. From a practical standpoint, daily data is more readily available, easier to clean, and sufficient for evaluating medium-term trading strategies such as those operating on a day-ahead forecast horizon. High-frequency data would require significantly more computational resources and robust infrastructure for storage and processing, which is beyond the scope of this academic project.
From a modeling perspective, daily data aligns well with our trading assumptions, in which positions are opened in the morning and closed at the end of the day, making it a natural choice for the objectives of this study.

2.2. Statistical Tests

Before proceeding with forecasting models, it is essential to assess the statistical properties of our time series data, particularly stationarity and causality.

2.2.1. Stationarity Test

Time series models such as ARIMA assume that the underlying data is stationary — meaning its statistical properties do not change over time. To test for stationarity, we used the Augmented Dickey-Fuller (ADF) test on each of our variables: the NIFTY 50, the Nikkei 225, and the VIX.

The null hypothesis of the ADF test is that the time series has a unit root (i.e., it is non-stationary). A small p-value (typically less than 0.05) allows us to reject the null hypothesis and conclude that the series is stationary. The test is applied to both the level and the first-difference of each series. As commonly observed in financial time series, the price levels are non-stationary, while their first differences tend to be stationary. Therefore, we will use the first difference of the Nikkei as an exogenous variable for prediction. The VIX, on the other hand, is already stationary without differencing.

2.2.2. Granger Causality Test

To determine whether the Nikkei 225 and the VIX indices provide predictive information about the NIFTY 50 index, we performe Granger causality tests. This test evaluates whether past values of one time series contain information that helps predict another series.

To this end, we test whether the Nikkei Granger-causes the NIFTY 50 returns. A low p-value (typically below 0.05) allows us to reject the null hypothesis that the exogenous variable does not Granger-cause the target series. The results will be presented in the forecasting section. These preliminary statistical tests provide a solid foundation for our modeling approach by validating both the need for differencing (in univariate models) and the relevance of using the Nikkei and VIX as informative variables.

2.3. Methodology

As explained previously, our methodology is structured around time series forecasting using econometric models enhanced by exogenous variables. The central objective is to generate reliable daily forecasts for the NIFTY 50 index that can be leveraged in a trading strategy (which will be explained in part IV).
We opted for SARIMAX models, as they are well-suited for capturing the autocorrelation structure of financial time series while allowing the integration of external information. Specifically, we explore four different configurations:

- **Univariate (Naive):** ARIMA model without any exogenous input.
- **Exogenous VIX only:** SARIMAX model using the VIX index as an external regressor.
- **Exogenous Nikkei only:** SARIMAX model using the Nikkei 225 index.
- **Combined VIX & Nikkei:** SARIMAX with both variables as exogenous inputs

To account for changing market dynamics, we implemente a **rolling-window re-estimation** strategy. At each step in the test set, we performe the following:

1. Updated the endogenous and exogenous datasets with the latest observed values.
2. Forecasted the next value using the current model.
3. Every 60 steps (approximately 3 months of trading days), re-estimated the optimal SARIMAX order using the `auto_arima()` function, minimizing the AIC criterion.

All models were initialized with an order of (0,1,0), corresponding to a random walk — and adjusted dynamically. We also applied a stability condition: if an order of (2,0,0) was selected, it was reverted to (0,1,0) to avoid numerical instability.
Our evaluation metrics include:

- **MAE** (Mean Absolute Error)
- **MSE** (Mean Squared Error)
- **CDC** (Correct Directional Change), which is particularly important in trading contexts.

This methodology allows us to assess not only the predictive accuracy of each model but also its practical utility for algorithmic trading decisions. Among all the evaluation metrics, the Correct Directional Change (CDC) score will be of primary importance, as it directly aligns with the trading strategy implemented in Part IV.

3. Forecasting

To forecast the Nifty 50, we employed two different approaches:

- A univariate approach using a classical ARIMA model;
- A multivariate approach incorporating exogenous variables.

3.1. Univariate forecasting

To forecast the Nifty 50, we adopted the following method: we initially fitted the model on the training data, and subsequently, for each new prediction, we updated the model parameters using a rolling window comprising the previous two months of data. The choice of a two-month rolling window is purely heuristic. Experimentally, the size of the rolling window does not significantly impact the forecast performance (at least if it is not too short).

The fitted orders was obtained by Grid-Search minimizing the AIC of several models. As we can expect, in a majority amount of case the bast model found was an ARIMA(0,1,0) (ie a random walk) :

$$y_t = y_{t-1} + \epsilon_t$$

Indeed, according to the arbitrage theory in financial markets, stock prices follow a martingale process. As a result, it is consistent to find that a random walk fits the data best, since the best forecast for the price of tomorrow is simply the price of today.

3.2. Multivariate forecasting using the Nikkei index

The main issue with the previous "naive" forecast is that it fails to capture any meaningful behavior or dynamics of the index. Essentially, it amounts to "gambling" the direction the index will take the next day. This means that, on average, we can expect to predict the correct direction only about 50% of the time (theorically even less). In addition, when developing a trading algorithm, the accuracy of direction prediction is a critical metric, often more important than standard error measures like MAE or MSE, as highlighted in [2]. It is why we developped a new forecasting method using an exogenous variable.

We considered using another index as an exogenous variable, provided it satisfies the following key conditions:

- Significantly correlated with the NIFTY 50
- Has market hours that closely align with those of the Indian market
- Sufficiently liquid to ensure reliable data

We chose Japan's main index, the Nikkei 225, as it satisfies all three conditions. In fact, the last two are clearly met: the Nikkei 225 is highly liquid, and its market opens at 5:30 AM (India time), well before the NIFTY 50, which opens at 9:15 AM (India time).

We will from now consider, as the exogenous variable, not the Nikkei 225 index level itself, but rather the difference between its previous day's close and the close on the day for which we aim to make the forecast in order to capture the overnight movement.
To ensure that there is a causality effect between this new exogenous variable and the index, we performed a Granger causality test :

Test	**Statistic**	**p-value**
la	6.6904	0.0098
χ^2 test	6.7043	0.0096
Likelihood test	6.6889	0.0097

Table 1: Granger Test result (lag=1)

The p-value is small enough to reject the null hypothesis which is here that the exogenous variable does not cause the Nifty index.

We now apply the same forecasting methodology as previously described, but incorporating the new exogenous variable. Once again, the most frequently selected model was an ARIMAX(0,1,0) :

$$y_t = y_{t-1} + \beta x_t + \epsilon_t$$

Where x_t is the differenc between the close of the $t-1$ day and the open of today t. This influence can also be observed in the values of the Beta coefficients :

Figure 1: Value of the β coefficient with respect to time

Let's remark that the Beta coefficient is particularly high during the COVID-19 stress period, which may indicate that the correlation between the two indexes was indeed quite strong.

3.3. Comparaison of the forecasts

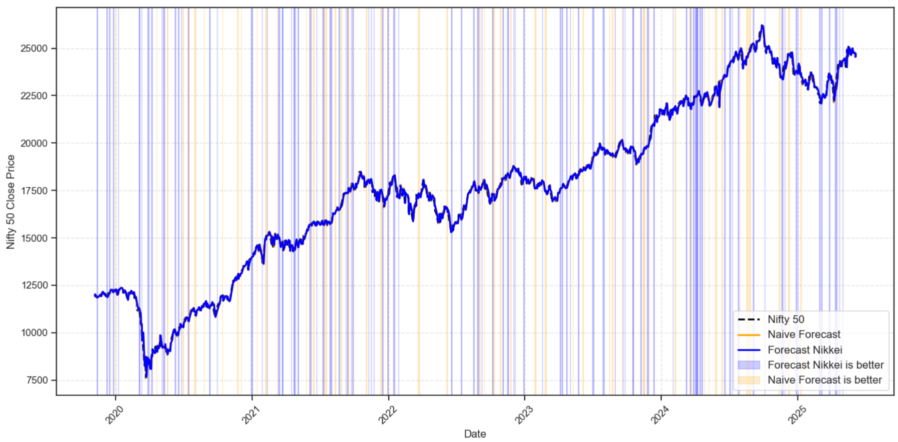

Figure 2: Directional Predictive Capacity of the Naive and Multivariate Forecasts

In the Figure 2 we plotted the two forecasts alongside the Nifty 50 index. As mentioned earlier, we are now focusing exclusively on the CDC metric. In the graph, the blue strips correspond to dates when the Nikkei forecast correctly predicted the direction of the index, while the naive forecast did not. Conversely, the orange strips represent dates when the naive forecast correctly predicted the direction, but the Nikkei forecast did not.

We can clearly observe that the Nikkei forecast outperforms the naive forecast in terms of CDC, especially during the specific period shown in Figure 3.

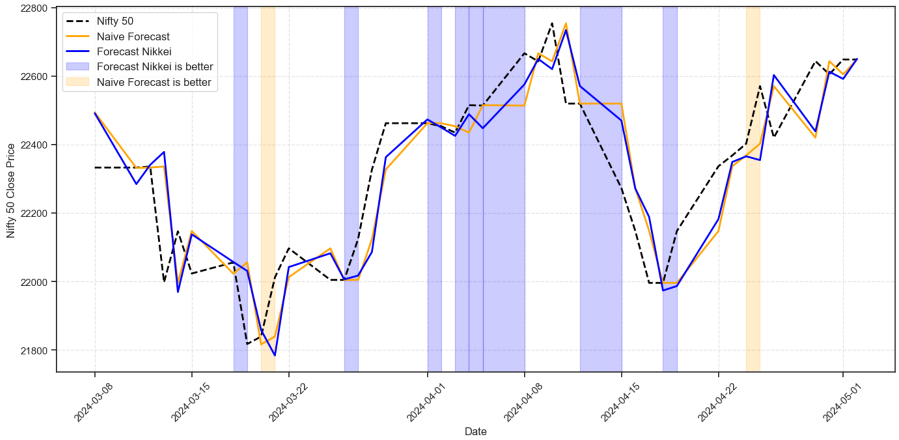

Figure 3: CDC of the Naive and Multivariate Forecasts (between the 2024-03-08 and 2024-05-02)

Over the entire test period, the CDC of the multivariate forecast is indeed higher than that of the univariate forecast as reported in this table :

Model	MAE	MSE	**CDC**
Naive Forecast	119	29535	**50.3%**
Nikkei Forecast	119	29112	**54.0%**

Table 2: Metrics of the forecasts

3.4. Multivariate forecasting using the VIX index

We now aim to enhance our forecasting model by incorporating the VIX index value observed at market open. Intuitively, we consider this value may help capture market sentiment and volatility expectations before the trading session begins, potentially improving the directional accuracy of our forecasts.
Here are the following results founded by the prediction :

Model	MAE	MSE	**CDC**
Naive Forecast	119	29535	**50.3%**
VIX Forecast	118	29756	**51.9%**
VIX + Nikkei Forecast	119	29346	**53.8%**
Nikkei Forecast	119	29112	**54.0%**

Table 3: Forecasting metrics using VIX as exogenous variable

Based on the CDC score, we find that the best-performing model is the one using only the Nikkei index as an exogenous variable. While the VIX alone also improves the forecast compared to the naive model, combining it with the Nikkei slightly reduces the overall performance. This suggests that the inclusion of both variables may introduce noise or redundancy rather than complementary information. We then prefer to use only the model with the Nikkei. We will then only focus on this method for the trading part.

4. Trading Algorithm

The final objective of this project is to capitalize on this forecast! A trading algorithm takes conditional decisions based on the information provided by our forecast and adjusts the position of the book adequately. Using this kind of tool obviously brings some risks tothe portfolio. That's why we will judge on its capacity to outperform the Nifty which represents a hold position. Our metric will be the return.

We do a serie of assumptions :

Assumption 1:
The liquidity is considered infinite and the price unique

Assumption 2:
We suppose constant interest rates

Assumption 3:
We invest on the index the morning just after the Asian open and we close the position at the end of the day

Assumption 4:
We consider negligible the trading costs as we only do two transactions per day at the maximum

4.1. One direction algorithm

To start, we consider an algorithm only allowed to buy the index. The concept is simple and is based on the following strategy :

$$\begin{cases} \frac{F_t - F_{t-1}}{F_{t-1}} > 0 : \text{We take a Long position with 100\% of the notional} \\ \frac{F_t - F_{t-1}}{F_{t-1}} \leq 0 : \text{We stay outside of the market} \end{cases}$$

Obviously, this basic algorithm isn't concluding and we decide to improve it using interest rates. In fact, rather then letting the position sleep in the portfolio, it is more interesting to make it grow on a risk free position when the forecast doesn't give any buying signal. We consider a rate equal to **4.3%** based on the current value of the Fed Fund according to [1].

Our algorithm becomes :

$$\begin{cases} \frac{F_t - F_{t-1}}{F_{t-1}} > 0 : \text{We take a Long position with 100\% of the notional} \\ \frac{F_t - F_{t-1}}{F_{t-1}} \leq 0 : \text{We invest our money on the risk free rate} \end{cases}$$

The results are already better as our portfolio grows when we don't have any positive signal from our forecast. However, we notice that even when the expected return, according to the forecast, is between 0 and the risk free rate, we are taking the risk to be long on the index. It means, we are taking a risk to win a forecasted return lower than the risk free return. To change that, we impose a new treshold before buying the index : the forecasted daily return should be higher than the daily risk free rate.

We obtain our final one direction trading algorithm :

$$\begin{cases} \frac{F_t - F_{t-1}}{F_{t-1}} > \frac{r_f}{252} : \text{We take a Long position with 100\% of the notional} \\ \frac{F_t - F_{t-1}}{F_{t-1}} \leq \frac{r_f}{252} : \text{We invest our money on the risk free rate} \end{cases}$$

In order to properly compare the different returns without any bias from the date, we consider the average return on more than 10,000 random time intervals.

Model	Average return (6 months)	Std Deviation
Hold	8.2%	12.5%
Only Long trading	8.6%	8.9%
Only Long trading with Nikkei	10.3%	8.0%

Table 4: One direction algorithm results

First of all, we notice that the standard deviation of our returns seems to decrease with the trading algorithm whereas the returns are improved. Both represent positive points in favor of our trading method. Finally, we can underline an outperformance from the Nikkei forecast in line with the previous CDC metric.

4.2. Two directions algorithm

Thanks to the previous section, we managed to improve our returns with one direction algorithm. We want now to improve it combining both long and short strategies. The following algorithm is applied :

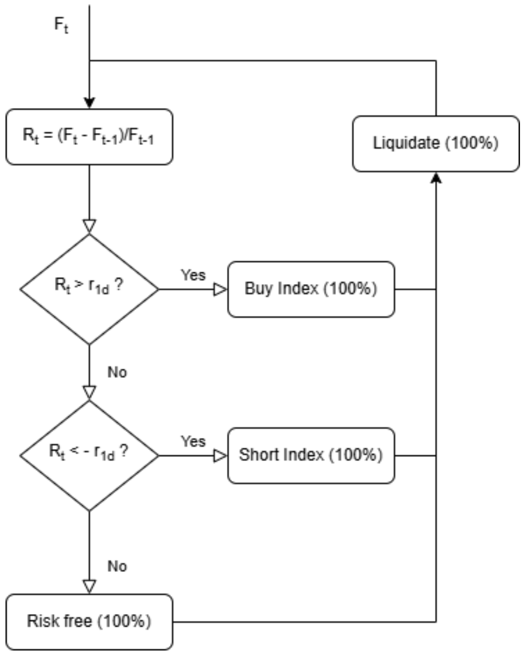

Figure 4: Long & Short Trading Algorithm

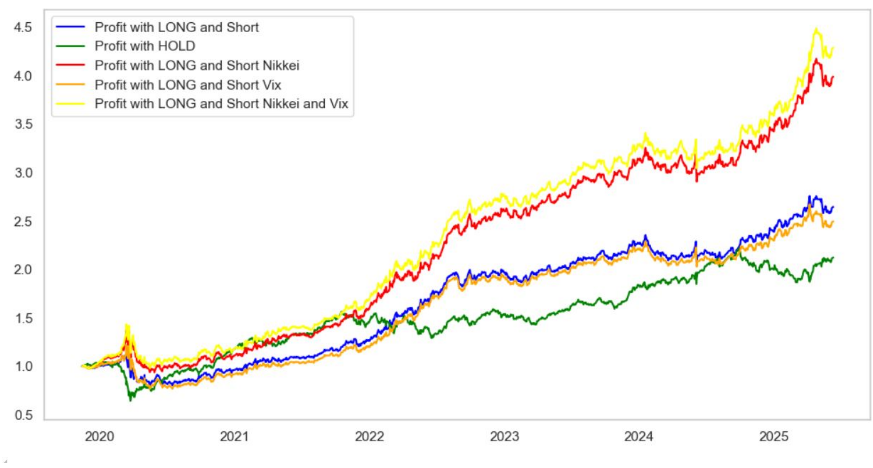

Figure 5: Long & Short Trading Algorithm

Interestingly, we observe that the combination of the Nikkei and the VIX generates higher profits than using the Nikkei alone. However, we cannot conclude that this first method is superior, as doing so would introduce hindsight bias. Despite the higher profits, the second approach may still be more robust it's possible that the improved performance of the first was simply due to chance over this specific time period.

Algorithm	Average return (6 months)	Std Deviation
Hold	8.2%	12.5 %
Only Long	10.3%	8.0%
Long & Short	13.0%	11.3%

Table 5: Long & Short algorithm results

The results are better in terms of returns with a double-direction algorithm but it increases the volatility of the returns. Moreover, we can underline the difficulties of our forecasted trading strategy to optimize the returns after the crisis and despite a very good profit during the period itself. The reason for the underperformance comes from the training set based on crisis data which bias our forecast.
Finally, despite difficulties post-crisis, our long-short algorithm is very interesting during the crisis itself. During March 2020, the return reaches 30%, where the index lost up to 35%. We also notice that every algorithms outperformed the index during this period.

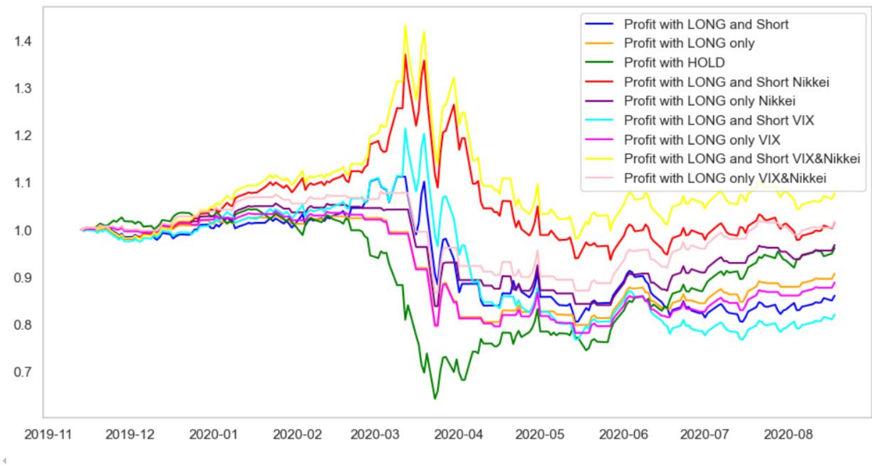

Figure 6: Trading results during Covid crisis

5. Conclusion

This project examines the use of SARIMAX models to forecast the NIFTY 50 index and implement trading strategies. As a main result, we find that incorporating the Nikkei 225 as an exogenous variable significantly improved directional forecast accuracy, while the VIX alone added value but reduced performance when combined with the Nikkei. Trading algorithms based on these forecasts outperformes a passive strategy, especially during volatile periods. Overall, the results highlight the potential of combining econometrics and financial intuition for market prediction and decision-making. As an opening, we could try to explore volatlity foercasting as presented in class by M. Targa.

References

[1] Fed fund. `https://fred.stlouisfed.org/series/FEDFUNDS`, 2025. Seen on June 8th 2025.

[2] Chris Brooks, A. G. Rew, and S. Ritson. A trading strategy based on the leadlag relationship between the spot index and futures contract for the ftse 100. *International Journal of Forecasting*, 17:[16–17], 2001.