Project report

Stock Price Prediction of CAC 40

Academic year 2023 - 2024

Authors :
Abdoul-Aziz TOURE
Dilane Desire KEUBOU KENFACK
Fiona Karlie FONKOU DAHVE
Isaac GANIYU

Supervised by :
Youssef ESSTAFA

December 2023

Acronyms and Abbreviations

ARIMA : Autoregressive Integrated Moving Average

ARIMAX : Autoregressive Integrated Moving Average with Exogenous Variables

ARCH : Autoregressive Conditional Heteroskedasticity

CAC : Cotation Assistée en Continu

CNN : Convolutional Neural Network

FARIMA : Fractionally Autoregressive Integrated Moving Average

GARCH : Generalized Autoregressive Conditional Heteroskedasticity

LSTM : Long Short-Term Memory

GDP : Gross Domestic Product

RNN : Recurrent Neural Network

RMSE :Root Mean Square Error

Table des matières

Table des figures

Liste des tableaux

Abstract

This study examines the evolution of the CAC40 index, a key indicator of the French economy. The study takes a multi-stage approach, incorporating descriptive statistics, methodological choices for model construction and in-depth analysis of our CAC40 index using each methodology. An initial descriptive phase has enabled us to understand the general trend in the series over recent years, paying particular attention to the fall in 2020 attributed to the global health crisis, COVID-19. In particular, the study discerns the stationarity of the return series, a crucial prerequisite for modelling time series.

Methodologically, in this study we select and apply a range of models, including ARIMA, FARIMA, ARIMAX, GARCH, XGBoost and LSTM. The ARIMA model seems to be significative but regarding the RMSE on the test set which is greater than 5000 and regarding it's prediction we can observe that it is not a good model to be used to predict CAC40 prices. The ARIMAX model stands out by incorporating lagged GDP growth, revealing a statistically significant impact on the CAC40 quaterly return. With an RMSE of 0.2% it seems to be a good approach to include macro-economic factors in stock return index prediction. The FARIMA model, capable of capturing both short- and long-term memory components, performs pretty well and is capable of capturing little volatility in our return series with an RMSE of 1.19% on our test set.

The GARCH model was used to effectively manage the challenge of conditional return volatility and performed relatively well on the test set with an RMSE of 1.18%. Meanwhile the LSTM model, despite impressive performance, exhibits a time lag in predictions of the returns, which may hinder its responsiveness to rapid changes in the index. Notably, the LSTM and GARCH(2,1) models excels in forecasting CAC40 returns, with low root mean square errors (RMSE) on both training and test sets. However, for the LSTM model, challenges emerge in predicting future values, as evidenced by a notable time lag and an RMSE of 1.37% in 20-day forecasts.

The LSTM model was equally used directly to forecast the CAC40 index prices but it doesn't perform well to predict future values with an RMSE of 283.10. In addition to all this methodologies, we equally performed an XGBoost model but it does not fits very well without using the actual values for new predictions because it lacks extrapolation.

Introduction

In the ever-evolving landscape of financial markets, the ability to anticipate and predict stock price movements is a paramount pursuit for investors and market analysts alike. The dynamic nature of stock prices, influenced by a multitude of factors ranging from economic indicators to geopolitical events, makes this task challenging yet crucial for informed decision-making. One notable index that reflects the performance of major companies and plays a significant role in the French financial market is the **CAC 40**. Established on 31 December 1987, in the aftermath of the October 1987 market crash, the CAC 40 index represents the 40 largest and most actively traded stocks on Euronext Paris (*www.lafinancepourtous.com*). Its creation marked a pivotal moment, responding to the changing landscape of stock market trading.

This study focuses on the development of predictive models for stock prices, leveraging historical data to enhance our understanding of market dynamics. The CAC 40, as a key indicator of the French economy, serves as a noteworthy case study.

The predictive models considered includes traditional time series analysis techniques such as ARIMA (Autoregressive Integrated Moving Average) models, volatility forecasting using GARCH (Generalised Autoregressive Conditional Heterosedasticity) and more contemporary approaches such as LSTM (Long Term Short Term Memory) in the field of machine learning. Acknowledging the multifaceted nature of market influences, we incorporate macroeconomic factors to our analysis, such as GDP growth, through the application of ARIMAX models.

Through this study, we aim not only to develop accurate stock price prediction models but also to contribute to the broader understanding of the dynamics influencing the CAC 40 index and, by extension, the French financial market. This study is structured into five key segments : commencing with a succinct overview of the description of the CAC 40's and it's evolution, followed by a brief exploration of methodologies employed by researchers in modeling and predicting the CAC 40. Subsequently, we present the methodologies used for our predictions, detailing the results obtained through various methods, after which we will evaluate each of our models, present their predictions alongside with the actual stock prices to assess the model's accuracy. The final step involves a thorough comparison of predictions generated by these different methods.

1 Presentation and description of the evolution of the CAC40

1.1 General overview

1.1.1 Presentation of CAC40

Created on 31 December 1987 by the "Compagnie des Agents de Change", **the CAC 40 index** which stands for Cotation Assistée en Continu is the main stock market index of the Paris Stock exchange. It was created following the crash of October 1987 which changed the monopole on stock market trading, but it was officially launched on 15 June 1988.

The CAC 40 is precisely a floating stock market index weighted according to the market capitalisation of companies. It reflects the performance of the 40 largest and most actively traded stocks among the 100 (one hundred) most traded companies on Euronext Paris (part of Euronext, Europe's leading stock exchange). It is quoted every working day from 9 a.m. to 5.30 p.m. and its value is updated every 15 seconds and published in real time on Euronext.

The CAC40 index is considered as a key indicator of the French economy, as it is representative of the Paris financial market in terms of both floating capitalisation and trading volumes. The companies in the CAC 40 are also representative of the different business sectors and reflect the overall economic trend of major French companies. The list is regularly reviewed to ensure that it remains representative.

1.1.2 Calculation and Composition of the CAC 40

The list of CAC 40 companies is updated quarterly by a committee of experts known as the Conseil Scientifique des Indices (CSI). If a company is no longer listed on the market, it is replaced in principle by one of the companies in the CAC Next 20 that meets the financial requirements for listing in the CAC 40 (liquidity of the stock, sufficient market capitalisation, significant daily trading in the stock, etc.).

Each of the 40 companies included in the index has a specific weighting, determined according to the value of the shares it has available on the market. These weights are adjusted for each company according to its floating capitalisation. Unlike other indices such as the German DAX, the CAC 40 is calculated on the basis of dividends not reinvested. Since the 1st December 2003, it has followed the standards of the main world indices, taking into account not only the total market capitalisation of companies, but also their free float, i.e. the proportion of shares actually available for purchase on the market.
The current composition of the CAC 40 in September 2023 is as follows :

1.2 Descriptive statistics

1.2.1 Data description

The CAC40 data used in this study were downloaded from the following website Yahoo Finance for the period 01-03-1990 to 27-10-2023. These data contain 04 variables, namely :

— Open : The opening price of the stock at the beginning of the trading day.

— High : The highest price the stock reached during the trading day.

— Low : The lowest price the stock reached during the trading day.

— Close : The closing price of the stock at the end of the trading day.

— Adj Close : The adjusted [1] closing price.

For the next part of the study, we used the variable **Close**.

The database therefore contains **8694 daily observations** of the close variable. However, 147 observations (less than 1.7%) are marked "Null", this corresponds to public holidays in France such as May 1st, Christmas, New Year [2] where the stock markets are closed. These observations have been removed from the database, giving us a database of 8547 daily observations.

1.2.2 Analysis of CAC40 price movements

This table below shows the central tendency and dispersion of the CAC40 closing price series.

Table 1 – Descriptive Indicators for the close price of CAC40

Descriptive indicators	CAC 40
Mean	4079
Median	4114
Maximum	7577.00
Minimum	1441
Standard Deviation	1488.98
Number of observations	8547

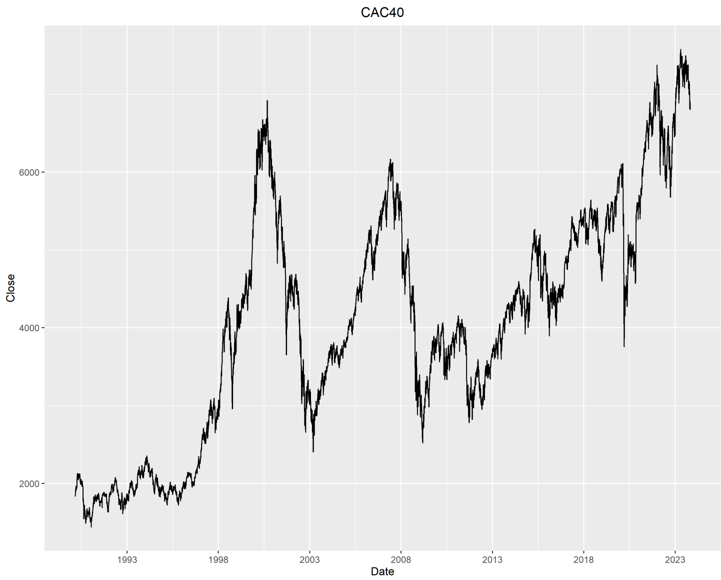
Source : Authors

Since the CAC 40 price series does not appear to be stationary, we cannot statistically interpret the trend parameters of this series. We will now focus on the evolution of the series. The graphic below shows the evolution of the CAC40's prices over the period under review :

1. This price is adjusted for events such as dividends, stock splits, or new stock offerings that occurred after the market has closed

2. when they fall during the week, because the stock markets are closed at weekends

FIGURE 1 – Evolution of CAC40's prices between 01-03-1990 and 27-10-2023



An analysis of the evolution of CAC40's prices between 1990 and 2023 shows that the index has seen some remarkable highs and lows. In 2000, encouraged by intense speculation in the telecoms, media and technology sectors, the index climbed to an all-time high of 7,000 points. However, it gradually collapsed and fell to almost 2,500 points in March 2003, the lowest level since 1997, due to a general decline in the markets in Europe and the United States.

After a remarkable upturn, the CAC 40 surpassed 6,000 points during 2007, reaching almost 6,100 points in June of that year. However, in January 2008, the index fell drastically, reaching 3,200 points at the end of the same year as a result of the subprime crisis, which affected the world's stock markets.

In 2017, the CAC 40 began to include companies from the technology sector, which had long been under-represented. Companies such as Atos, Dassault Systèmes and, in 2019, Thalès have all been included, leading to a significant change in the index, which was previously dominated by the financial and industrial sectors.
2019 was a remarkable year for the CAC 40, with a record payout of €60.2 billion, including €49.2 billion in dividends and €10.7 billion in share buybacks. However, in 2020, the CAC 40 suffered one of the biggest falls in its history, losing 12.28% as a result of the global health crisis, the Covid19.

In addition to the above economic analyses, figure 1 shows that the CAC40 price series does not appear to be stationary. So, in practice, we can use the returns of the series.

1.2.3 Analysis of CAC40 Returns

The return of a stock can be calculated by the formula :

$$R_t = \frac{P_t - P_{t-1}}{P_{t-1}}$$

where : R_t is the return at time t ; P_t is the price at time t ; P_{t-1} is the price at time t-1 .

For the CAC40, the table below shows the central tendency and dispersion of the series of returns between 1990 and 2023.

Table 2 – Descriptive Indicators for the returns of CAC40

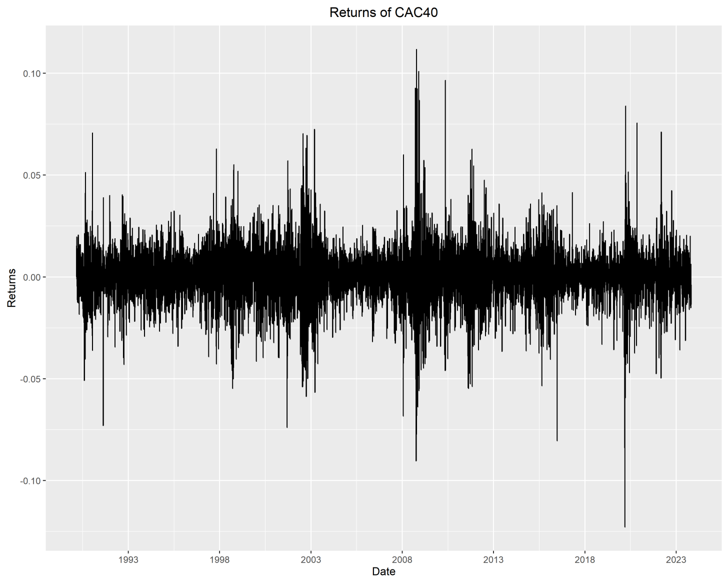
Descriptive indicators	CAC 40
Mean	0.0002458
Median	0.0004308
Maximum	0.1117617
Minimum	-0.1227677
Standard Deviation	0.01358704
Number of observations	8546

Source : Authors

The average return on the CAC 40 over the last thirty years is 0.025%. The highest return is 11.17%, while the lowest observed is -12.27%. This return corresponds to that of the year 2020 due to the global health crisis COVID 19, when the value of the CAC 40 experienced one of the biggest falls in its history.

The graphic below shows the evolution of the CAC40's return over the period under review :

Figure 2 – Evolution of CAC40's return between 02-03-1990 and 27-10-2023



This series appears to be stationary, which makes it easier to model, unlike the closing price series. This stationarity is confirmed by the ADF and KPSS tests appended to the document (graphic 20)

2 Literature review and methodology

2.1 Literature review

In this study, our objective is to predict the value of the CAC40 or its return. Several similar studies have been carried out. In this section we list some of the studies we have found on the subject. The aim of this literature review is to identify clues for the treatment of our data, the methodology to be adopted in our study and the models to be used.

Ariyo et al. (2014) presents extensive process of building stock price predictive model using the ARIMA model. Published stock data obtained from New York Stock Exchange (NYSE) and Nigeria Stock Exchange (NSE) are used with stock price predictive model developed. Results obtained revealed that the ARIMA model has a strong potential for short-term prediction and can compete favourably with existing techniques for stock price prediction. The variable "close" was chosen for the study because it reflects activity throughout the trading day.

Hennani and de Séverac (2015) apply the finite mixture of ARMA-GARCH model instead of AR or ARMA models to compare with support vector machines (SVM) in forecasting financial time series. They analyze three stock market indices S&P 500 from New York Stock Exchange, FTSE 100 from London Stock Exchange and CAC 40 from Paris Stock Exchange with daily returns data. The data cover the period between 01 July 1996 and 01 July 2013. To improve the forecasting ability and make it possible, they transform this data into daily returns.

Elliot and Hsu (2017) present four models to predict the stock price using the S&P 500 index as input time series data : a martingale model and ordinary linear models used as baselines, a generalized linear model and a LSTM model. They then compare the performance of the models.

Selvin et al. (2017) use three different deep learning architectures for the price prediction of NSE (National Stock Exchange of India) listed companies. The deep learning architectures used were RNN (Recurrent Neural Network), LSTM (Long Short Term Memory) and CNN (Convolutional Neural Network). They then compare the performance of the three algorithms.

2.2 Methodology

For this study, the variable to be predicted is the "close" variable. This variable is suitable because it reflects what has happened on the market during the day. It is also the most widely used variable in the literature. First, we delete the missing values of this variable because they correspond to public holidays. Next, we divide the database into training and test sets, with 80% of our data in the training set and 20% in the test set.

Throughout this study, we will be setting up several models with the aim of modeling and predicting the value of the CAC40 or its return. We will then make a comparative analysis of the models obtained by measuring the performance of our models. To do this, we will use the test sample. The measure of performance we will use is the Root Mean Square Error (RMSE) which is defined as follows :

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{1}$$

with N the number of samples in the corresponding data set, y_i the i-th actual value and $\hat{y}_i$ the i-th predicted value

3 Modelling the time series

3.1 ARIMA Model

Autoregressive integrated moving average (ARIMA) models are stationary series differentiated models with stochastic trend or stochastic seasonality. They are obtained by differencing the series in a given order such that the differentiated series becomes stationnary. Differentiation helps to correct a non-stationary series with deterministic components such as a trend or seasonality

3.1.1 Theory and model building

A stationary process X_t admits a minimal ARIMA(p,d,q) representation if it satisfies :

$$\Phi(L)(1-L)^d X_t = \Theta(L)\varepsilon_t, \quad \forall t \in Z \tag{2}$$

where

$$\begin{cases} \Phi(L) = I - \phi_1 L - \phi_2 L^2 - \ldots - \phi_p L^p \\ \Theta(L) = I + \theta_1 L + \theta_2 L^2 + \ldots + \theta_q L^q \end{cases} \tag{3}$$

with the following conditions :

1. $\phi_p \neq 0$ et $\theta_q \neq 0$
2. The polynomials Φ and Θ, of respective degrees p et q, have no common roots and their roots are of module > 1
3. ε_t is a white noise of variance σ^2

To estimate the parameters of an ARIMA model, we proceed by using the Box and Jenkins approach. This approach consists of 3 majors steps. Firstly we have to , then we have to chose the

- Choice of the differentiation order such that the series obtained after differentiation is stationary then choice of p and q which are the orders of the AR and MA components respectively.
- Estimation of the parameters ϕ_i and θ_j of the AR and MA components respectively.*
- Finally we have to evaluate the significance of the parameters and validate the white noise assumption for the residuals.

To model an ARIMA series, the orders p and q of the AR and MA models are obtained by looking at the autocorrelations and partial autocorrelations. Very often, these orders are not obvious. In such cases, we can obtain upper bounds for p and q and then select a model by minimising a penalised criterion of the AIC or BIC type.

Once the order and type of model have been chosen, several approaches are possible for estimating the model coefficients. We can use the Yule-Walker equations or estimate the parameters using the maximum likelihood or conditional least squares method. Once the model has been estimated, it is important to validate or not our modelling choices. To do this, we proceed in the following two phases :

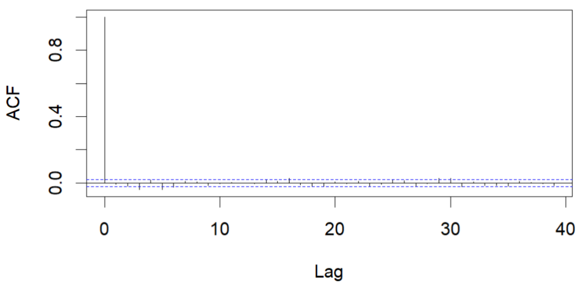
- **Residue analysis :** Once an ARIMA model has been estimated, we can obtain an estimate of ϵ_t. We then test the whiteness of these residuals using the Box-Pierce test or the Ljung-Box test.

— **Parameter significance test :** The idea here is to compare nested ARMA(p,q) models to see whether it makes sense to reduce or increase the p and q orders.

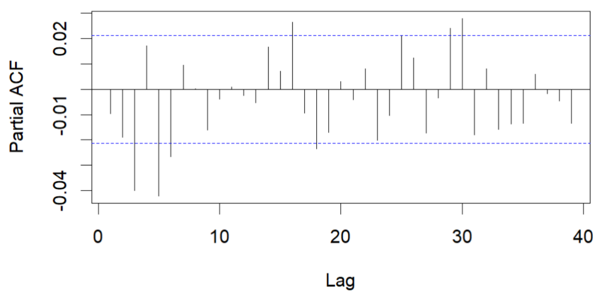
3.1.2 Results

We chose to model our price series using an ARIMA model. In order to verify the type of ARIMA process that our series represents, we begin with a graphical analysis of the series to check whether there is any trend or seasonality in our series. After our first graphical analysis, we realise that there is a trend in our series. To correct this, we chose to differentiate the series at the first order. Once the series had been differentiated, we realised from a graphical analysis that there is no more trend in our series of prices and that the series appeared to be stationary. To check this statistically we carry out a Dickey Fuller test on the differentiated series. The results of the test shows that the differentiated series is stationary. We plotted the ACF and PACF of our differentiated series to determine the orders of our ARIMA. The ACF and PACF graphs are shown below.

FIGURE 3 – ACF et PACF of the Return series



ACF of the integrated series



PACF of the integrated series

A first analysis of the ACF might lead us to believe that our integrated series is a white noise. However, if we look at the PACF we see that it does not cancel out after a certain rank. It therefore seems that our series is indeed generated by an ARIMA(p,d,q) process. To determine the orders p and q, we will use the Box and Jenkins method. We will choose their maximum values, then construct the arima model for each combination and choose the best model based on the AIC criterion. The table of results for each combination is given in appendix 8. Based on these results, we choose ARIMA(3,1,3) because it minimises the AIC. Once the orders have been chosen, we proceed to an analysis of the residuals of our model. The graph in appendix 21 gives us an overview of the residuals of our ARIMA model. We also performed a test to assess whether the autocorrelations of the errors were significant or not. The p-value of the Box-Pierce test is 0.952, so we can't reject the null hypothesis that the residuals are not autocorrelated. The model can the be udes and interpreted. Nevertheless, the RMSE calculated on our test set was very large (5906.06) meaning that it's performance is not pretty good. The model is not good for prediction, as we can see from the graph in the appendix 18, where the prediction on our model was quiet constant compared to the real values of the differentiated series.

3.2 FARIMA Model

Fractionally differenced autoregressive integrated moving average models (FARIMA) are considered here because they account for both the short- and long-memory components that are present in many financial processes. FARIMA models allows the differencing order d to be fractionnal.

3.2.1 Theory and model building

Let us assume that a time series X_t has a mean of zero. Using the Box and Jenkins notation, the general form of the FARIMA (p, d, q) model can be defined as the stationary solution of

$$(1 - B)^d \Phi_p(B)(1 - \Theta_q(B))X_t = \theta(B)\varepsilon_t, \tag{4}$$

where B is the backward shift operator (i.e., $BX_t = X_{t-1}$), $\Phi_p(B)$ is the p-order autoregressive polynomial, $\Theta_q(B)$ is the q-order moving average polynomial, and ε_t is the error or noise term, which is assumed to be uncorrelated and with zero mean, and d is the differencing order, allowed to be fractional. The d value is related to the Hurst parameter H by the relationship d=H- 0.5. The range of interest is between 0 and 0.5 ; when d = 0, reduces to the classical ARMA model. The higher the value of d, the higher the intensity of long memory displayed by the model. When d is greater than 0.5, the model is nonstationary.

In this section, we estimate a FARIMA model on the series of CAC40 returns. We use the function *arfima* from the package *forecast* in R. With this function, an ARFIMA(p,d,q) model is selected and estimated automatically using the Hyndman-Khandakar (2008) algorithm to select p and q and the Haslett and Raftery (1989) algorithm to estimate the parameters including d.

3.2.2 Results

After estimating our model on the training sample, we can summarise the estimated parameters values in the following table :

Parameter	Estimated value
d	$4.58 * 10^{-5}$
ma1	0.01131407
ma2	0.03534799
ma3	0.04903497
ma4	-0.01448262
ma5	0.05469293

TABLE 3 – Estimated parameters

After this estimate, we move on to validating the model. The estimated residuals are as follows :

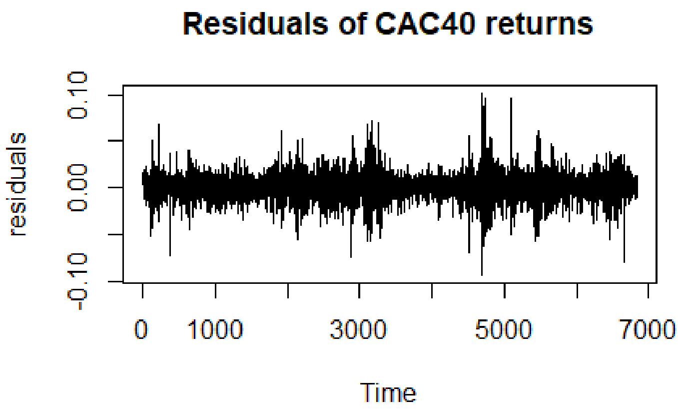
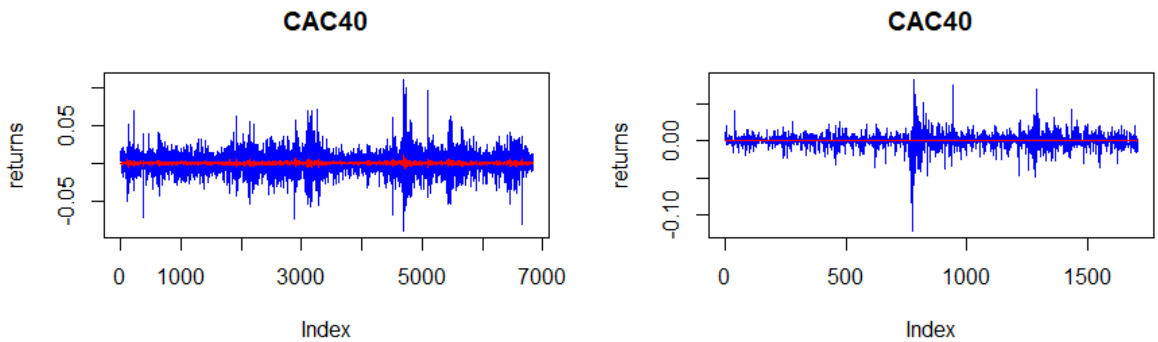


FIGURE 4 – Residuals of CAC40 returns estimated with FARIMA

The p-value of the Box-Pierce test is **0.955**. We can not reject the null hypothesis that the residuals are not autocorrelated. The model can therefore be used and interpreted.

We are now interested in the performance of this model. On the train set, we obtained a RMSE of **1.39%** and on the test set, we obtained a RMSE of **1.19%**. The predictions can be represented.



Train set

Test set

FIGURE 5 – Performance of the FARIMA model

3.3 ARIMAX Model

The ARIMAX model is an extension of the Box–Jenkins autoregressive-moving average (ARIMA) model with explanatory exogenous variables (X). In ARIMAX models we include exogenous variables which are likely to have an impact on the series of interest.

3.3.1 Theory and model building

Before modelling the ARIMAX series, we need to be sure that the Y_t and X_t series are stationary. The general equation of an ARIMAX can be written as :

$$Y_t = a_0 + \Phi(L)Y_{t-1} + B(L)X_t + \Theta(L)\varepsilon_t \tag{5}$$

where :$\Phi(L)$corresponds to the AR component of Y_t

$\Theta(L)$ corresponds to the MA component ofY_t

$B(L)$ is called the TRANSFER FUNCTION

$\Phi(L)Y_t = \Theta(L)\varepsilon_t + D(L)X$
$Y_t = \sum_{i=1}^{p} \phi_i Y_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \sum_{i=1}^{b} z_{t-d}$

To model an ARIMAX series, we proceed in the following steps : First, using the Box and Jenkins method, we determine the orders p, d and q of our Y_t series. Once we have determined these orders, we need to determine the order of the polynomial B(L), i.e. which lags of the exogenous variable X_t are correlated with the series Y_t. To do this, we use the cross-correlation function. This function is defined by :

$$\hat{\rho}_{xy}(h) = \frac{\hat{\gamma}_{xy}(h)}{\sqrt{\hat{\gamma}_x(0)\hat{\gamma}_y(0)}} \tag{6}$$

where $\hat{\gamma}_{xy}(-h) = \hat{\gamma}_{yx}(h)$ determines the function for negative lags and

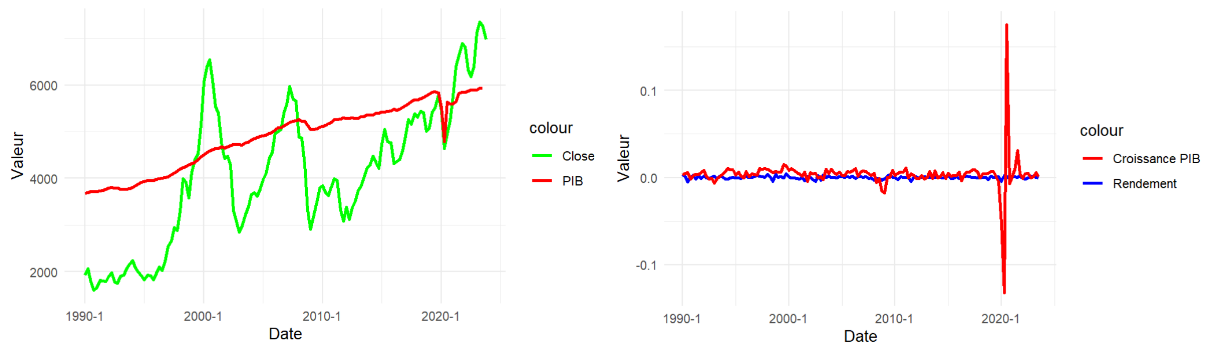
$$\gamma_{\hat{xy}}(h) = \frac{1}{n}\sum_{t=1}^{n-h}(x_{t+h} - \bar{x})(y_t - \bar{y}) \tag{7}$$

The sample cross-correlation function can be examined graphically as a function of lag h to search for leading or lagging relations in the data. Once all these parameters have been estimated, we can now build our ARIMAX model. In literature there are several methods for training this type of model, and each of these methods depends on the characteristics of the Y_t series. In our case, we will use the **arima** function implemented in R, including the exogenous variables, to train our model.

3.3.2 Results

For our ARIMAX model, we are more interested in modelling the return using the GDP growth rate. This choice is based on literature review, but also because of the stability of the two series and because the result will be easy to interpret. Also, we want to work directly on a stationary series and on series which have the same magnitude. Given that GDP values are given in quarterly values, we begin by changing the values of our returns to quarterly values by aggregating and taking the average. Below we have a representation of the evolution of the two series. On the graph on the left, we can see that the two series are evolving in the same direction, i.e. they both have an increasing trend. The return and GDP growth rate series on the right appear to be stationary.

FIGURE 6 – Evolution of the GDP series and the stock series



GDP and closing price GDP growth and Returns

Once the order of the ARIMA orders of the return series has been determined, to build our ARIMAX model, we only need to include the exogenous variables to see if they improve the prediction. GDP values are taken from the INSEE database.

Before modelling our ARIMAX, we checked that our series are stationary through a Dickey Fuller test. Once this was verified, we determined the order of the return series ARIMA process using the Box and Jenkins procedure as explained above. We chose an ARIMA(3,0,3) for our return series. Once the order was determined, we checked which lags of GDP growth was correlated with our return series by using the cross-correlation function. The graph of the cross-correlation is given in appendixe 22. We observe that there is a peak at the first lag, which may suggest that there is a significant correlation between the return in year t and GDP growth in year t-1. We modelled the return series with the lag of GDP growth and obtained the following results :

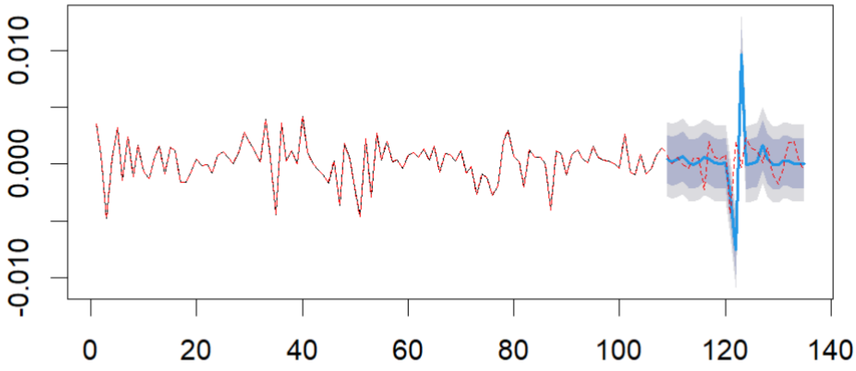
Coefficient	Estimate	Std. Error	z value	Pr(> $\|z\|$)
ar1	-9.4746e-01	1.0232e-01	-9.2600	< 2.2e-16 ***
ar2	-1.0039e+00	4.1610e-02	-24.1259	< 2.2e-16 ***
ar3	-9.1305e-01	9.8423e-02	-9.2768	< 2.2e-16 ***
ma1	8.4804e-01	1.3025e-01	6.5109	7.468e-11 ***
ma2	9.3342e-01	8.2975e-02	11.2494	< 2.2e-16 ***
ma3	8.3982e-01	1.3123e-01	6.3998	1.556e-10 ***
intercept	1.1659e-05	2.1172e-04	0.0551	0.95609
GDP_{t-1}	5.5395e-02	3.1291e-02	1.7703	0.07667 .

TABLE 4 – ARIMAX Model Coefficients and Standard Errors

We can see that the coefficient associates to the GDP is significant at the threshold of 10%. This suggests that the lagged GDP growth has a significant impact on the value of the return at that given threshold.

After analysing the residuals of our ARIMAX model, whose representation can be found in figure 19, we can see that the residuals are autocorrelated and normally distributed (this results had also been verified using statistical tests). The RMSE of our model is of approximately 0.0016 on the train set which suggest that it performs better than the ARIMA model with an RMSE of 52 on the train set (11). We made a prediction based on our test data of the GDP growth to see if our model is capable of predicting return variations. The following graph illustrates the prediction of our model.

FIGURE 7 – Returns Forecast using GDP growth



It seems as our model is capable of predicting some variations in the stock returns. However these variations doesn't perfectly match the variations of the stock returns.

3.4 GARCH Model

The GARCH (Autoregressive Conditional Heteroscedasticity) model plays a key role in describing financial series, given the heteroscedastic behaviour of their variance. This has always been poorly taken into account by ARIMA models, in which variance is assumed to be unconditional with respect to time.

3.4.1 Theory and model building

We say that (X_t) is a conditionally heteroscedastic generalized autoregressive process if it can be written in the form :

$$\begin{cases} X_t = \sigma_t \varepsilon_t \\ \sigma_t^2 = \omega + \alpha_1 X_{t-1}^2 + \ldots + \alpha_p X_{t-p}^2 + \beta_1 \sigma_{t-1}^2 + \ldots + \beta_q \sigma_{t-q}^2 \end{cases}$$

where (ε_t) is a white noise of variance 1, $\omega > 0$ and for all i, $\alpha_i \geq 0$ et $\beta_i \geq 0$. On note $(X_t) \sim \text{GARCH}(p, q)$.
The process X_t is stationary when $\sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} \beta_j < 1$, in this case $(X_t^2) \sim \text{ARMA}(max(p, q), q)$

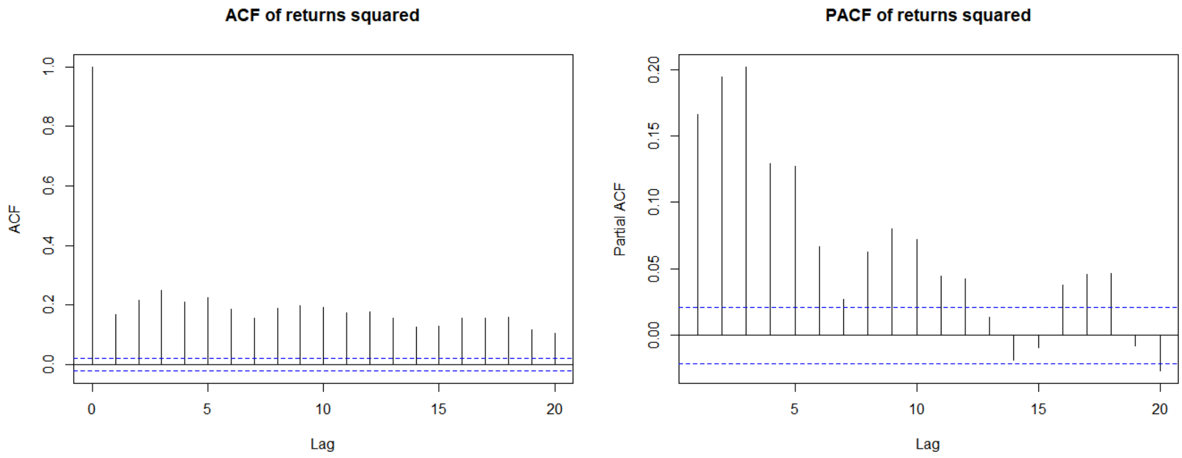
A stationary GARCH(p,q) process is therefore centred and not autocorrelated, and its variance is :

$$\gamma_0 = \frac{\omega}{1 - \alpha_1 - \ldots - \alpha_p - \beta_1 - \ldots - \beta_q}$$

NB : when $q = 0$, this is known as the ARCH(p) model.
A GARCH can be modelled in 04 steps :

1. checking for the existence and presence of ARCH effect This involves studying the volatility of the generating process and its stationarity. For the volatility, this can

Figure 8 – ACF and PACF of Returns squared



ACF of Returns squared

PACF of Returns squared

be checked by analysing the graph of the time series or make a Conditional heteroskedasticity test (ARCH test). This test is based on the significance of the regression, performed on the square of the time series studied.

2. Identification of Model Order (p, q)
   The aim here is to identify the maximum orders of the model, using the ACF (for q_max) and the PACF (for p_max) of **the square of the series**. However, if those two graphs do not provide plausible values for p and q, we can choose a set of possible pairs[3] of p and q and select the optimal pair, which is the one with the smallest information criterion (AIC, BIC).
3. Estimation of the parameters of the model
   The parameters ω, α_i and β_j are typically estimated using maximum likelihood estimation (MLE). The estimation is done by fitting the model to the historical data and finding the set of parameters that maximize the likelihood function.
4. Model validation
   Once the parameters have been estimated, a diagnostic must be carried out. This involves checking the whiteness of the residuals.

3.4.2 Results

First of all, we checked in our series the presence of an ARCH effect. This was done graphically, showing that the CAC40 return series exhibits conditional volatility (see graphic 2). This is confirmed by an ARCH test on the graphic 24. We can therefore model the returns series using a GARCH.
In order to determine the orders of the model, the graphs below show the ACF and PACF of the square of returns of CAC40 . Since the orders of p and q appear to be high, we set a maximum value of 3, as mentioned above. We then estimated various GARCH models, and the results obtained are appended.

3. $p, q \in \{1, \cdots, 3\}$ for example

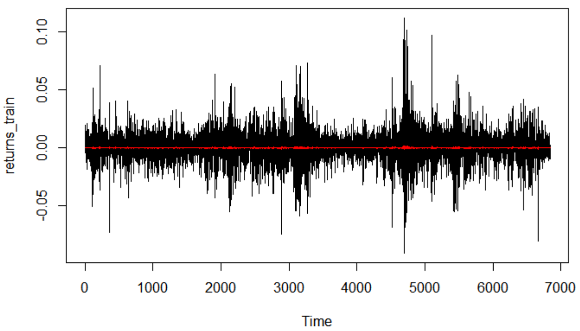
The model selected according to the AIC criterion is **GARCH (2,1)**. The coefficients of the model are shown below and are all significant at the 5% level :

FIGURE 9 – Coefficients of GARCH Model

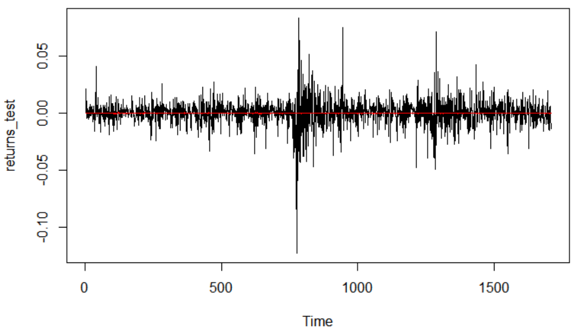
```
*---------------------------------*
*          GARCH Model Fit        *
*---------------------------------*

Conditional Variance Dynamics
-----------------------------------
GARCH Model      : sGARCH(2,1)
Mean Model       : ARFIMA(0,0,0)
Distribution     : norm

Optimal Parameters
------------------------------------
        Estimate  Std. Error  t value Pr(>|t|)
mu      0.000533    0.000131   4.0553 0.000050
omega   0.000004    0.000001   3.1716 0.001516
alpha1  0.054021    0.013174   4.1005 0.000041
alpha2  0.048695    0.016409   2.9675 0.003002
beta1   0.877979    0.011647  75.3804 0.000000
```

The model obtained is valid because : The residuals are non-autocorrelated (p value of Ljung box : $p \approx 0.4335$), normal and and have an ARCH effect (p value of LM test : $p \approx= 0$) . Those results are presented in appendixe (graphic 23).

According to the performance of the model, we separated the data into 80% for the train set and 20% for the test. So the train set corresponds to the data for the first 26 years of the series, and the test set to the last seven years. The results obtained on the train and test samples are good. In fact, the RMSE for the train sample was **1.37%** and for the test sample **1.18%**. The graphs below illustrate this :



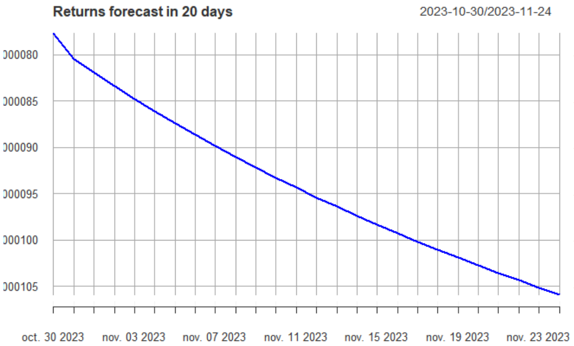
Model performance on the train set



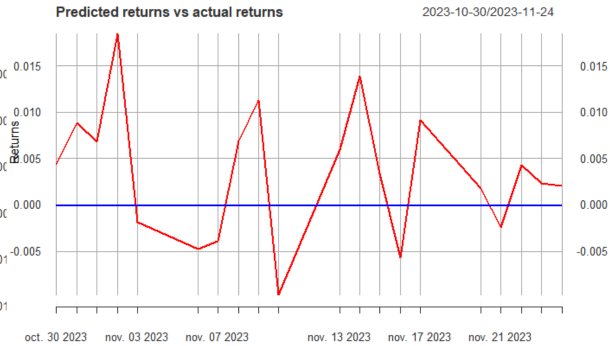
Model performance on the test set

FIGURE 10 – Performance of the GARCH model on the returns

A prediction of CAC40 returns was then made for the next 20 trading days of the series, starting on 30/10/2023. The result gives us an RMSE of **0.77%** and the graphs below show the predicted and actual return values.



Returns Forecast in 20 days



Comparison Returns Predicted and actual Returns

FIGURE 11 – Prediction of CAC40 Returns with the GARCH model

3.5 Xgboost model

3.5.1 Theory & modeling of Xgboost

XGBoost is a model based on trees. It stacks as many trees as desired, with each additional tree attempting to minimize the error. The overall concept is to combine several simple and weak predictors to construct a stronger one.

The most important formula in the XGBoost documentation [1] is how predictions are calculated. It is a fairly simple formula :

$$\hat{y}_i = \frac{1}{K}\sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathcal{F} \tag{8}$$

The prediction y_i is estimated, x_i represents a feature vector, $f_k(x_i)$ denotes the calculated values for each tree, and K is the total number of trees.

As you can observe, the XGBoost model works primarily like an additive model in relation to each individual tree. Let's examine f_k to understand the computation of tree scores and discern the specific type of function being referred to in this context.

$$f_t(x) = w_{q(x)}, w \in (R^T), \in q : (R^d) \longrightarrow \{1, 2, ..., T\}$$

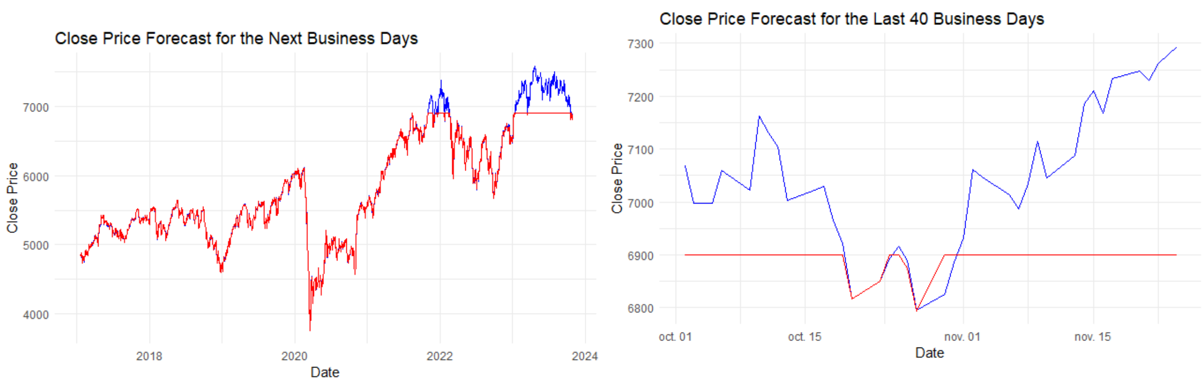
q(x) is a function that assigns features x to a specific leaf of the current tree t. $w_{q(x)}$ is then the score of the leaf for the current tree t and the current features x.

In essence, after training your model—arguably the most challenging aspect of the task—prediction becomes a matter of simply identifying the appropriate leaf for each tree based on the features and summing up the values assigned to each leaf.

We will perform this model on the price of CAC40.

3.5.2 Results & limit

Xgboost's model seems to perform pretty well on the test base, as the root mean square error is approximately equal to **135.68**. So, we can say that the model fits the test base data pretty well. However, here on the test set, we use the actual data to predict each new value. So we tried then to predict the future values, without using any actual data. We made it it over 20 working days. It seems that, the model perform less well, because the root mean square error is **246.09**.



Model performance on the test set

Close Price Forecast of the last 40 business days

FIGURE 12 – Performance of the Xgboost model

This difference between the error value on the test base and the predictions of 20 days is due to a limitation of the Xgboost model, which is the lack of extrapolation. The problem is that the predictions made by the Xgboost model are based on sums of the predictions of the different tree leaves. No transformation is applied to these values : no scaling, no log, no exponential, nothing. This means that XGBoost can only make a good prediction for situations already encountered in the training history. It can't capture trends !

3.6 LSTM model

3.6.1 Theory

LSTM stands for long short-term memory networks, used in the field of Deep Learning. LSTM is a Recurrent Neural Network (RNN) based architecture that is widely used in time series forecasting.

Recurrent neural networks are neural networks which have some way of remembering the previous values inputed into them, and output based on both the input, and the remembered value. This is done by three subprocesses : forget gate, input gate, and output gate.

The LSTM rectifies a huge issue that recurrent neural networks suffer from : short-memory. Using a series of 'gates,' each with its own RNN, the LSTM manages to keep, forget or ignore data points based on a probabilistic model. With repeated epochs, gradients become larger or smaller, and with each adjustment, it becomes easier for the network's gradients to compound in either direction. This compounding either makes the gradients way too large or way too small. While exploding and vanishing gradients are huge downsides of using traditional RNN's, LSTM architecture severely mitigates these issues.

3.6.2 Model building

For our architecture, we will use two LSTM layers followed by a dense layer of 1 node for output.

We have decided to use this architecture to model two quantities :

— the values of the CAC40

— the returns of the CAC40, whose expression is given below :

$$R_t = \frac{P_t - P_{t-1}}{P_{t-1}}$$

To train this model, we normalised the training base using a standardisation.

We have chosen a sequence value of 30. This means that a value will be predicted in our model by considering the 30 values that precede it. The value of this hyper-parameter was chosen on the basis of empirical tests. Our training data set of 6,837 samples enabled us to construct 6,807 sequences of 30 samples. These sequences will be used directly to train our model. Similarly, our test data set of 1,710 samples enabled us to construct 1,690 sequences of 30 samples. These sequences will be used directly to test our model.

The first layer of our architecture is an LSTM layer with 50 units and a "relu" activation function. This is followed by an identical layer. The last layer is a Dense layer with one unit. Dense implements the operation :

$output = activation(dot(input, kernel) + bias)$.

Kernel is a weights matrix created by the layer, and bias is a bias vector created by the layer.

The architecture of our model is therefore as follows :

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 lstm (LSTM)                 (None, 30, 50)            10400

 lstm_1 (LSTM)               (None, 50)                20200

 dense (Dense)               (None, 1)                 51

=================================================================
Total params: 30651 (119.73 KB)
Trainable params: 30651 (119.73 KB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

FIGURE 13 – Architecture of the LSTM model

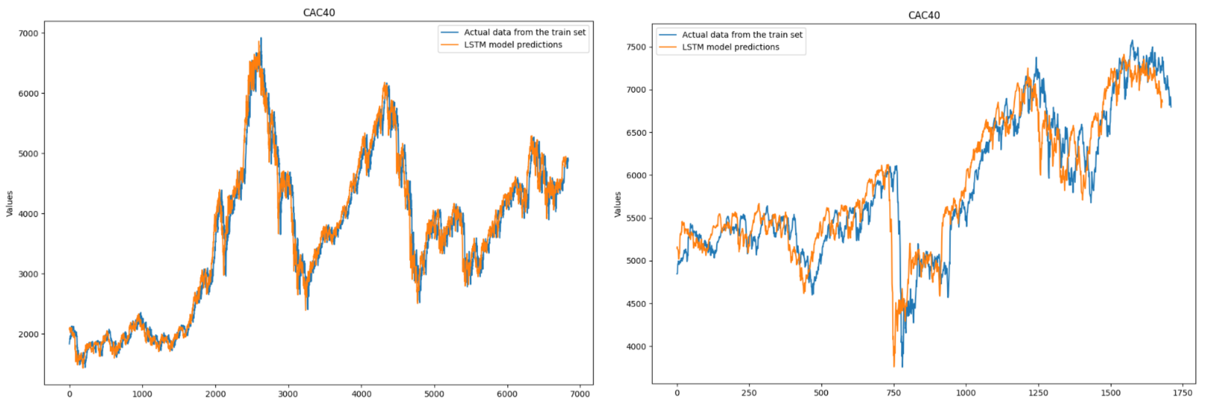
For this model, we will measure the performance on the test set. We will also predict the future 20-day values of the CAC40 index and its return. These are two different prediction approaches :

— On the test set : the data used for each prediction is real data. Although we predict a value at time t, the values used for the prediction at time t+1 are the true values observed before t+1 and available in the test set. We therefore make a prediction at horizon 1 each time.

— Future predictions : we now assume that the actual values are no longer observed, so once the prediction has been made at time t it is used for the prediction at time t+1. We therefore make a prediction with a 20-day horizon.

3.6.3 Results

- **CAC40 index**
  After training our model, we evaluate its performance, which at first glance seems quite impressive. We quantify the quality of our model and obtain an RMSE of **77.60** on the training set and **70.50** on the test set. We plot the model's predictions on the following graphs :



Model performance on the train set   Model performance on the test set

FIGURE 14 – Performance of the LSTM model

We note that the predicted values are very similar to the actual values. However, the predictions do not match the true values. There is a time lag between the two graphs in each data set. This means that our model incorporates the information contained in the real data with a delay.

We have also produced 20-day forecasts for the CAC40 index. We compare these forecasts with the corresponding values actually observed. This leads to a large RMSE since here, the next predictions is done with the previous one. The RMSE is **283.10**.

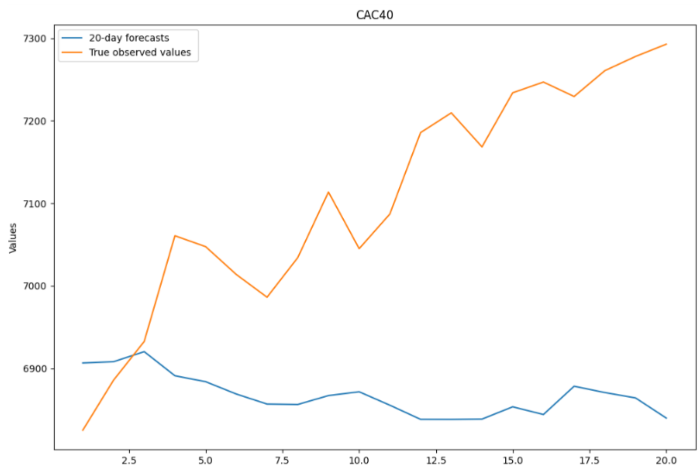
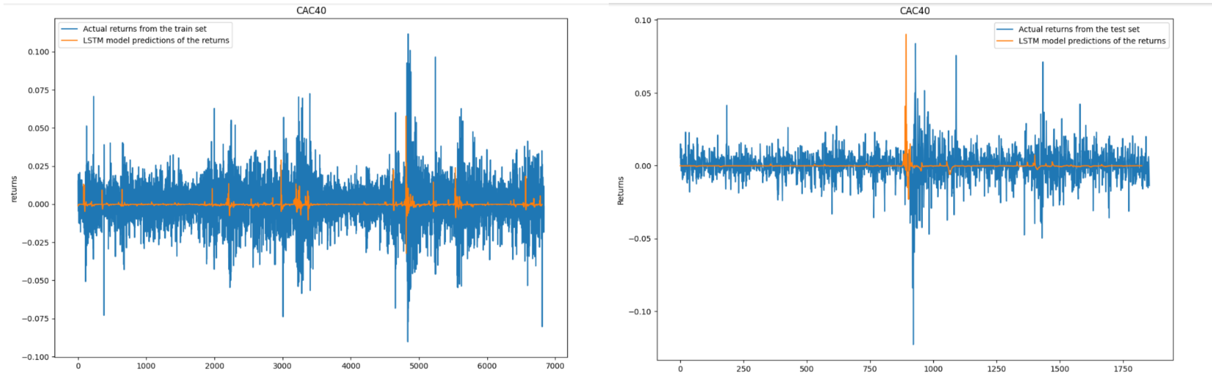


FIGURE 15 – 20-day forecasts

As we can see, the model failed to predict the rise in the value of the CAC40 after the fall observed in the test sample.

- **Returns of the CAC40**

  We are nos interested by the results of our model when we model the returns of the CAC40 instead of the values. As in the previous section, we present the performance of the model on the train and test sets :



Model performance on the train set   Model performance on the test set

FIGURE 16 – Performance of the LSTM model on the returns

We obtain an RMSE of **1.377%** on the training sample and **1.184%** on the test sample. We can furthermore notice that the model underestimates the volatility of the series and rarely predicts negative returns.

Now we can, as before, produce 20-day forecast of the returns and compare it to the true returns observed.

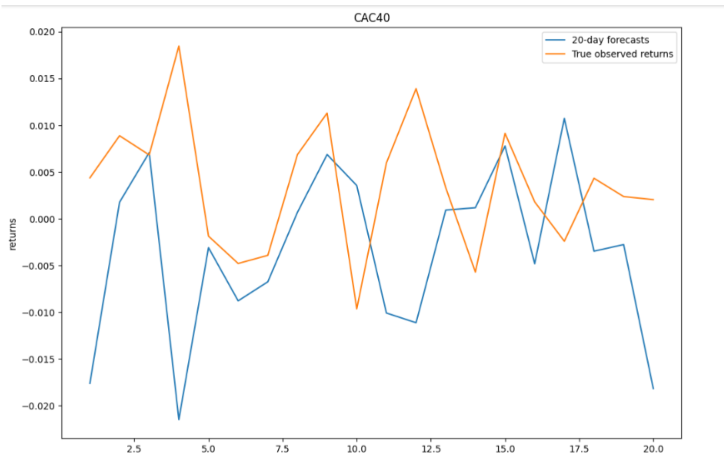


FIGURE 17 – 20-day forecasts of the returns

Predicted returns are fairly volatile over the 20 days. And the RMSE is : **1.369%**. Once again, there seems to be a lag between observed values and predictions.

4 Comparative analysis of models

Throughout this study, we have implemented several models. These models estimate different quantities, namely the value of the CAC40 or its return. We will now compare the models we have estimated. To do this, we separate the models into two groups : Those that estimate the value of the CAC40 and those that estimate the return of the CAC40. We summarise the performances in the following tables :

Models	RMSE TRAIN	RMSE TEST	RMSE 20-DAY FORECAST
ARIMA	52.064	5906.058	NA
LSTM	77.60	70.50	283.10
XGBOOST	5.13	135.68	246.10

Table 5 – Comparison of price's models

	RMSE TRAIN	RMSE TEST
GARCH	1.37%	1.18%
FARIMA	1.39%	1.19%
LSTM	1.377%	1.184%

Table 6 – Comparison of return's models

We note that the price models perform very differently. The ARIMA model is applied to the test set without using the real data, but provides largely unsatisfactory results. The LSTM and XGBOOST models are applied to the test set using the real data from the test base for each new prediction. In this context, the LSTM model seems better than the XGBOOST model. However, we note that the performance of these two models decreases as soon as the future is predicted, without using real data. It therefore seems more appropriate to estimate CAC40 returns rather than prices.

We note that the return's models perform very similarly to each other. The FARIMA and GARCH models make a prediction similar to a straight line. The LSTM model, on the other hand, makes fairly volatile predictions but with a greater prediction error of 1.369% when we predict without using the actual data.

We have produced a last model which is not similar to the others. It is the ARIMAX model, which estimates quarterly returns. We obtained with this model a RMSE of **0.293%** on the test set. This is much lower than the other RMSE obtained. However, these figures are not perfectly comparable because with ARIMAX we estimated quarterly returns.

Conclusion

The main objective of this study was to predict CAC 40 prices, based on several time series models. Using a dataset of CAC 40 from the period 01-03-1990 to 27-10-2023, we meticulously filtered and then performed descriptive analyses to familiarize with the data. Then, we explored two groups of models, the first group have 03 models : ARIMA, LSTM and XGBOOST for modelling prices. The second group : FARIMA, GARCH and LSTM models was used for stationarity purposes to model returns series. All those series were segmented into training set (80%) and test set (20%), to ensure the robustness of our predictive analysis.

Finally, we compared each group of models with the RMSE criteria : for price's models the **LSTM** was the best model with 70.50 as RMSE on test set ; On the other hand, the **GARCH(2,1)** for return's models was the best , since it has the smallest RMSE (1.18%)on the test set. The comparative analysis of the different methodologies highlighted the strengths and limitations inherent in each. As a prediction, we forecast returns over a 20-day horizon (between October 30, 2023 and November 16, 2023). The GARCH and LSTM models gave us a forecast error of 0.77%. Overall, Our study however come up against limitations, but tried to challenge the complexity of financial markets, where even sophisticated models can capture trends and patterns, but may fail to make ultra-precise predictions.

The ARIMAX model stands out by incorporating lagged GDP growth, revealing a statistically significant impact on the CAC40 quaterly return. With an RMSE of 0.293% it seems to be a good approach to include macro-economic factors in stock return index prediction since it seems like it is capable of capturing some variability in the return series.

Bibliography

Références

[1] Documentation XGBOOST. URL : `https://xgboost.readthedocs.io/en/latest/tutorials/model.html`.

— Selvin, S., Vinayakumar, R., Gopalakrishnan, E. A., Menon, V. K., & Soman, K. P. (2017, September). Stock price prediction using LSTM, RNN and CNN-sliding window model. In 2017 international conference on advances in computing, communications and informatics (icacci) (pp. 1643-1647). IEEE

— Alberto Montanari, Murad S. Taqqu & Renzo Rosso (1997). Fractionally differenced ARIMA models applied to hydrologic time series : Identification, estimation, and simulation.

— Ariyo, A. A., Adewumi, A. O., & Ayo, C. K. (2014, March). Stock price prediction using the ARIMA model. In 2014 UKSim-AMSS 16th international conference on computer modelling and simulation (pp. 106-112). IEEE.

— Elliot, A., & Hsu, C. H. (2017). Time series prediction : Predicting stock price. arXiv preprint arXiv :1710.05751.

— Hennani, N., de Séverac, M. B. Application of ARMA-GARCH model and Support Vector Regression in Financial Time Series Forecasting.

— Robert H. Shumway David S. Stoffer. Time series analysis and it's application with R examples Fourth edition.

Annexes

Table 7 – Weights of Companies in the CAC 40 Index as of September 30, 2023.

Company	Weight (%)	Company	Weight (%)
LVMH	11.02	STMICROELECTRONICS	1.60
TOTALENERGIES	9.54	DASSAULT SYSTEMES	1.45
SANOFI	7.08	LEGRAND	1.43
L'OREAL	5.81	MICHELIN	1.27
SCHNEIDER ELECTRIC	5.51	ORANGE	1.24
AIR LIQUIDE	5.13	SOCIETE GENERALE	1.04
AIRBUS	4.61	VEOLIA ENVIRON.	1.02
BNP PARIBAS ACT.A	4.34	PUBLICIS GROUPE SA	1.01
VINCI	3.42	EDENRED	0.91
HERMES INTL	3.36	THALES	0.86
AXA	3.31	CREDIT AGRICOLE	0.76
SAFRAN	3.31	ARCELORMITTAL SA	0.75
ESSILORLUXOTTICA	3.21	CARREFOUR	0.57
STELLANTIS NV	2.64	RENAULT	0.49
DANONE	2.06	TELEPERFORMANCE	0.44
PERNOD RICARD	1.98	ALSTOM	0.42
KERING	1.97	EUROFINS SCIENT.	0.41
CAPGEMINI	1.68	BOUYGUES	0.39
ENGIE	1.63	WORLDLINE	0.37
SAINT GOBAIN	1.61	UNIBAIL-RODAMCO-WE	0.34

Source : Euronext

Ordre ARIMA	p	d	q	AIC	BIC
ARIMA(0,0,0)	0	0	0	149145.54	149159.65
ARIMA(0,0,1)	0	0	1	137779.82	137800.98
ARIMA(0,0,2)	0	0	2	128302.94	128331.15
ARIMA(0,0,3)	0	0	3	121120.74	121156.01
ARIMA(0,1,0)	0	1	0	92814.41	92821.46
ARIMA(0,1,1)	0	1	1	92814.85	92828.96
ARIMA(0,1,2)	0	1	2	92814.85	92836.01
ARIMA(0,1,3)	0	1	3	92808.36	92836.57
ARIMA(1,0,0)	1	0	0	92834.49	92855.65
ARIMA(1,0,1)	1	0	1	92835.02	92863.23
ARIMA(1,0,3)	1	0	3	92829.05	92871.37
ARIMA(1,1,0)	1	1	0	92814.90	92829.00
ARIMA(1,1,1)	1	1	1	92816.87	92838.03
ARIMA(1,1,2)	1	1	2	92813.45	92841.66
ARIMA(1,1,3)	1	1	3	92809.65	92844.92
ARIMA(2,0,0)	2	0	0	92835.18	92863.40
ARIMA(2,0,1)	2	0	1	92831.84	92867.11
ARIMA(2,0,2)	2	0	2	92841.34	92883.66
ARIMA(2,1,0)	2	1	0	92815.03	92836.19
ARIMA2,1,1)	2	1	1	92811.24	92839.45
ARIMA(2,1,2)	2	1	2	92813.77	92849.04
ARIMA(2,1,3)	2	1	3	92811.47	92853.78
ARIMA(3,0,0)	3	0	0	92837.09	92872.36
ARIMA(3,0,1)	3	0	1	92837.88	92880.20
ARIMA(3,0,3)	3	0	3	92839.87	92896.29
ARIMA(3,1,0)	3	1	0	92808.92	92837.13
ARIMA(3,1,1)	3	1	1	92810.48	92845.74
ARIMA(3,1,2)	3	1	2	92812.04	92854.36
ARIMA(3,1,3)	3	1	3	**92804.48**	**92853.85**

TABLE 8 – ARIMA results table with AIC and BIC

FIGURE 18 – Prevision of the ARIMA model

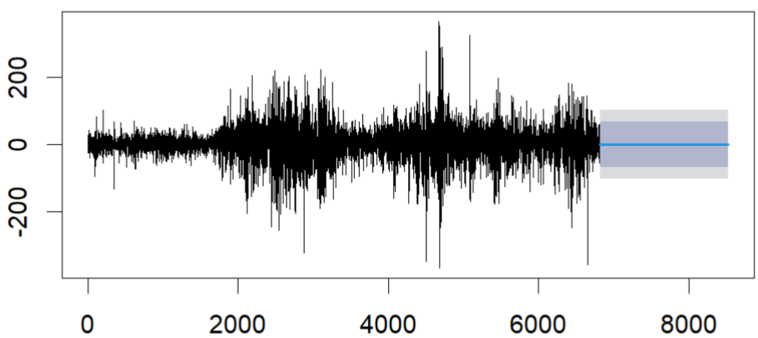


FIGURE 19 – Stationarity results of Returns

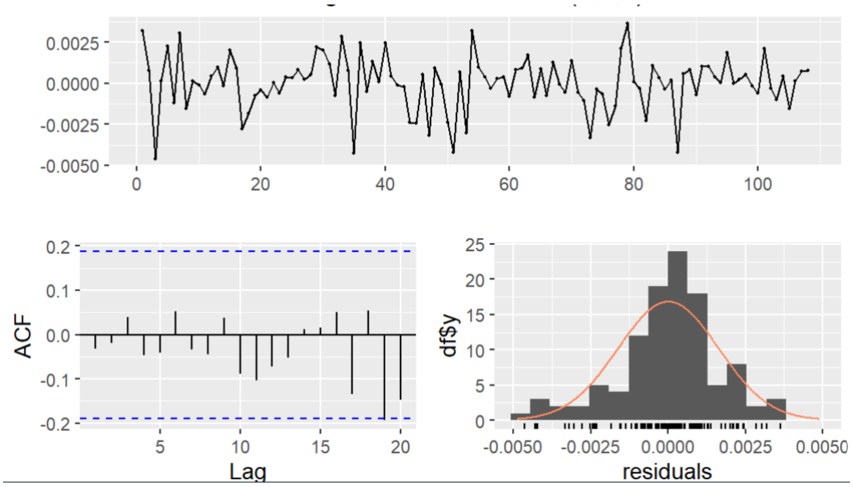


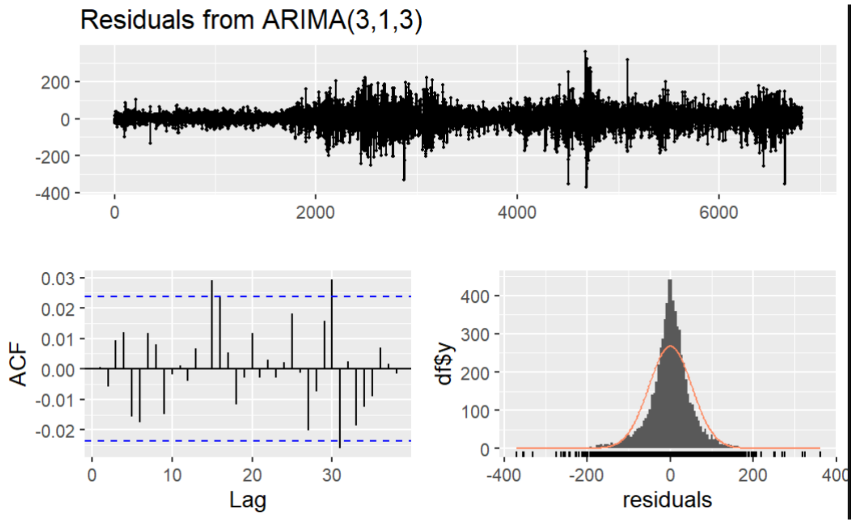
FIGURE 20 – Stationarity results of Returns

```
Avis : p-value smaller than printed p-value
        Augmented Dickey-Fuller Test

data:  returns
Dickey-Fuller = -20.915, Lag order = 20, p-value = 0.01
alternative hypothesis: stationary


#######################
# KPSS Unit Root Test #
#######################

Test is of type: mu with 12 lags.

Value of test-statistic is: 0.0491

Critical value for a significance level of:
                10pct  5pct 2.5pct  1pct
critical values 0.347 0.463  0.574 0.739
```

GARCH (p,q)	Akaike	Bayes	Shibata	Hannan-Quinn
1 1	-5.958916	-5.954921	-5.958917	-5.957538
1 2	-5.958577	-5.953583	-5.958579	-5.956854
1 3	-5.958291	-5.952298	-5.958293	-5.956224
2 1	**-5.959787**	-5.954793	-5.959788	-5.958064
2 2	-5.959597	-5.953604	-5.959599	-5.957530
2 3	-5.959545	-5.952553	-5.959547	-5.957133
3 1	-5.959467	-5.953474	-5.959468	-5.957399
3 2	-5.959276	-5.952284	-5.959278	-5.956864
3 3	-5.959470	-5.951479	-5.959473	-5.956713

TABLE 9 – Table of GARCH models with AIC and BIC

FIGURE 21 – Residual analysis for the ARIMA(3,1,3) model



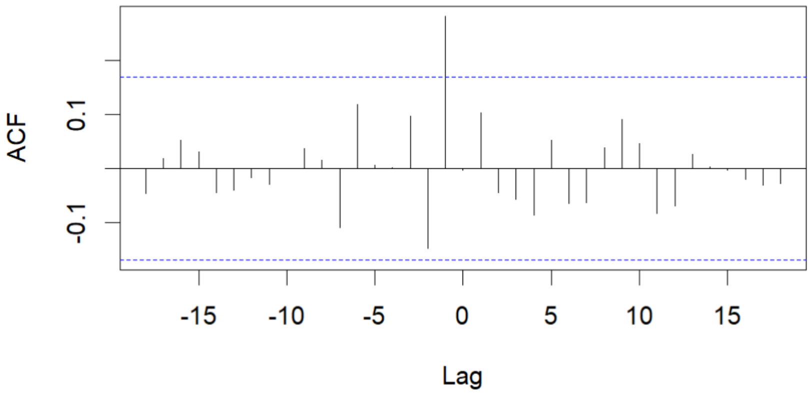
	ar1	ar2	ar3	ma1	ma2	ma3
Coefficient	-0.193	0.5358	-0.0465	-0.8264	-0.7427	0.5692

TABLE 10 – ARIMA Model Coefficients and Standard Errors

	ME	RMSE	MAE
Training set	-0.001904287	52.06439	36.13548

TABLE 11 – Training Set Error Measures

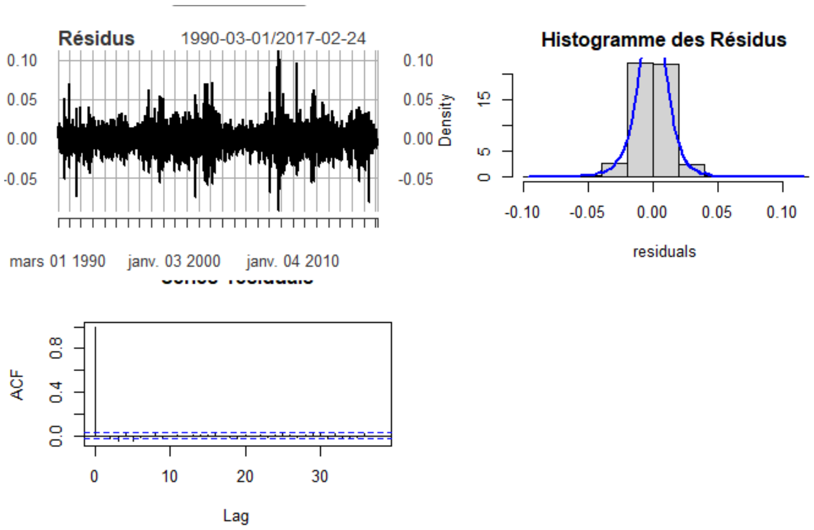
FIGURE 22 – Cross-correlation between the Return series and the GDP Growth series



Error Measures/Model Statistics	Value
RMSE	0.00161942
MAE	0.001188694
MPE	90.2697
MAPE	144.4172
AIC	-1061.78
BIC	-1037.64

TABLE 12 – Results of the ARIMAX(3,0,3) model

FIGURE 23 – Residual analysis for GARCH(2,1) model




```
        ARCH LM-test; Null hypothesis: no ARCH effects

data:  returns
Chi-squared = 1306.8, df = 12, p-value < 2.2e-16
```

```
        Box-Ljung test

data:  residuals
X-squared = 0.6133, df = 1, p-value = 0.4335

        ARCH LM-test; Null hypothesis: no ARCH effects

data:  residuals
Chi-squared = 1057.4, df = 12, p-value < 2.2e-16
```

FIGURE 24