Machine Learning & Predictive Analysis

Stock Prediction

Y.Hiraji, B.Hajji, Y.Lourarhi, A.Mansouri, R.Nazih

Dr. AHIDAR Adil
Ecole Centrale Casablanca

January 19, 2024

Table Of Contents

Problem Overview

Scientific Project

Problem Overview

Why Predict Stock?

- Maximize profits
- Predict the economy
- Implement suitable economic policies

Challenges

- Stochastic nature
- Multiple factors

Project Objectives

Scientific Project

What are the Goals?

- Build a working ARIMA (Autoregressive integrated Moving Average) model
- Build a working LSTM model
- Build a working feature fusion LSTM - CNN model
- Outputs: predicted daily closing for Apple (AAPL) (Application)
- Compare different results
- Find in which kind of stock these models can perform better

ARIMA Models

Scientific Project

Time Series

Example 1:

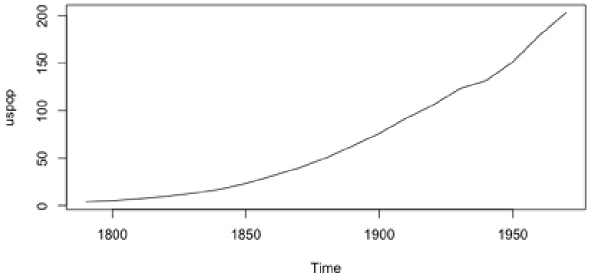


Figure: Population size of the USA between the years 1790 and 1990

Time Series

Example 2:

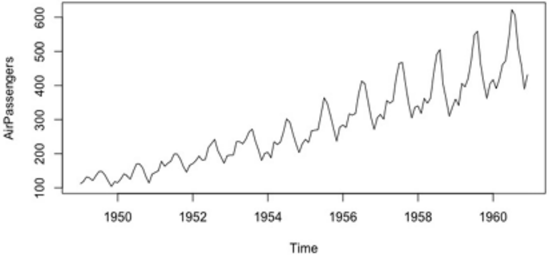


Figure: Monthly number of passengers (in thousands) between the years 1949 and 1960 in the airlines.

Time Series

What is it?

- Mathematically, a time series is a series of data indexed by time.

How does a time series decompose?

- A trend (T_t)
- A seasonality (S_t)
- A residual or error (Xt)

$$Y_t = T_t + s_t + X_t \tag{1}$$

Time Series

The trend

- The trend corresponds to an increasing or decreasing behavior of a series over time. It often reflects a long-term phenomenon of growth or decline.

The seasonality

- Seasonality reflects the presence of a periodic phenomenon that repeats itself throughout the time series.

The residual

- The residual of the model corresponds to the part of the time series that the decomposition does not explain. A time series cannot be entirely decomposed solely based on trend and seasonality.

ARIMA

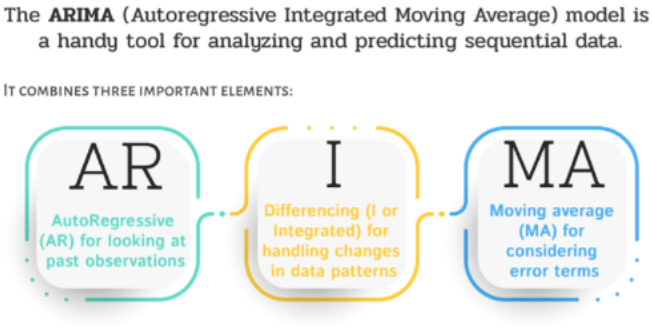


Figure: ARIMA Model (AR) + (I) + (MA)

ARIMA

The AR process

- Concretely, if we consider a stationary process X_t, we consider it to be autoregressive of order p if we can explain its value at time T using its previous p terms.
  Mathematically, this means that:

$$\forall t : X_t = \sum_{i=1}^{p} \alpha_i X_{t-i} + \varepsilon_t \tag{2}$$

ARIMA

The I

- The I in the ARIMA model stands for 'integrated.'
  It addresses the need for time series data to be stationary in order to be suitable for modeling. If a time series is not stationary, meaning its statistical properties change over time, we must often apply differencing.
  By differencing time series, it is possible to remove the trends they exhibit to make them stationary.

ARIMA

The MA process

- Let X_t be a time series, we consider it to be an MA (Moving Average) process of order q if we can express its value at time t as a linear combination of random errors (white noise).
  Mathematically, this is expressed as:

$$\forall t : X_t = \varepsilon_t + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i} \tag{3}$$

Box-Jenkins method

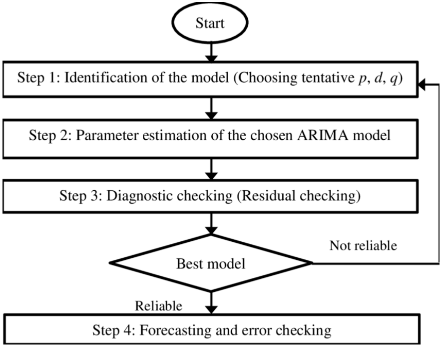
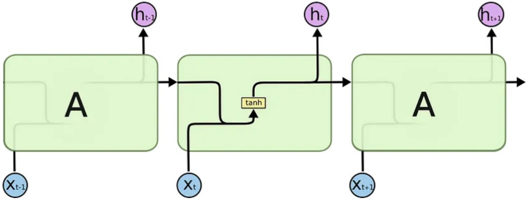


Figure: Forecasting procedure using Box-Jenkins approach

LSTM Model

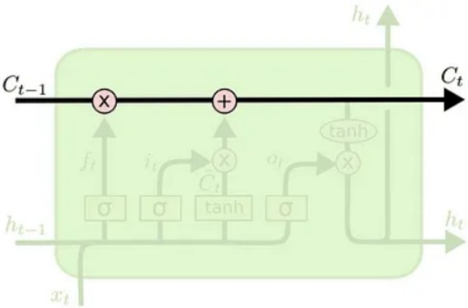
Scientific Project

RNN cellul



$$h_t = \tanh(W_{hh}h_{t-1} + W_{xh}x_t) \qquad (4)$$

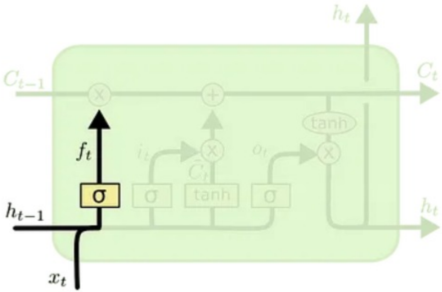
LSTM construction

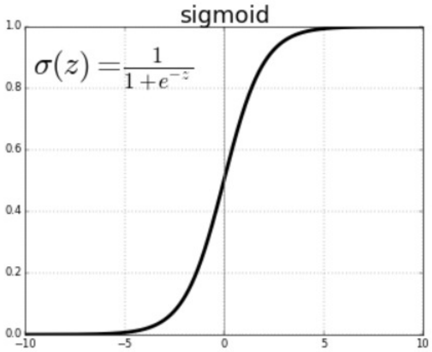


$$c_t = c_{t-1} + \tanh(W_{hh}h_{t-1} + W_{xh}x_t) \quad (5)$$

$$h_t = \tanh(c_t) \quad (6)$$

Forget Gate





Fonction d'activation sigmoïde

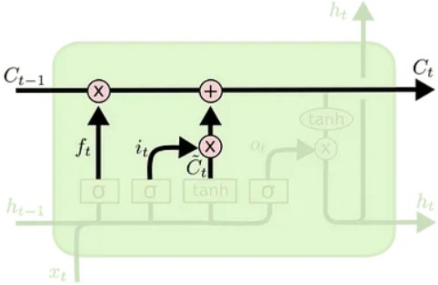
Forget Gate

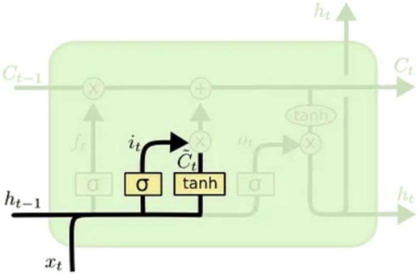
$$f_t = \sigma(W_{hf}h_{t-1} + W_{xf}x_t) \tag{7}$$

$$c_t = f_t * c_{t-1} + \tanh(W_{hc}h_{t-1} + W_{xc}x_t) \tag{8}$$

$$h_t = \tanh(c_t) \tag{9}$$

Input Gate





Input Gate

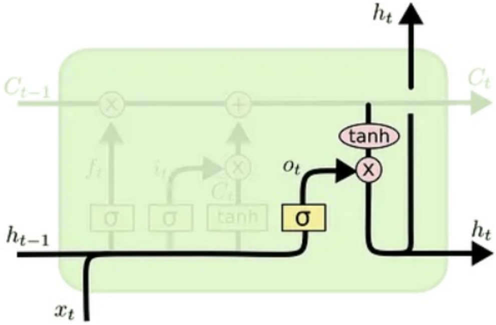
$$f_t = \sigma(W_{hf}h_{t-1} + W_{xf}x_t) \tag{10}$$

$$i_t = \sigma(W_{hi}h_{t-1} + W_{xi}x_t) \tag{11}$$

$$c_t = f_t * c_{t-1} + i_t * \tanh(W_{hh}h_{t-1} + W_{xh}x_t) \tag{12}$$

$$h_t = \tanh(c_t) \tag{13}$$

Output Gate



$$f_t = \sigma(W_{hf}h_{t-1} + W_{xf}x_t) \tag{14}$$

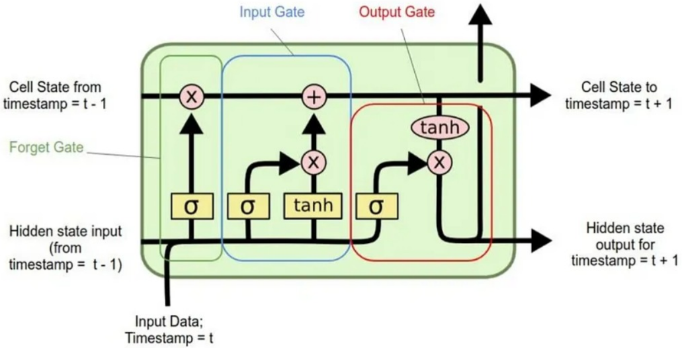
$$i_t = \sigma(W_{hi}h_{t-1} + W_{xi}x_t) \tag{15}$$

$$o_t = \sigma(W_{ho}h_{t-1} + W_{xo}x_t) \tag{16}$$

$$c_t = f_t * c_{t-1} + i_t * \tanh(W_{hh}h_{t-1} + W_{xh}x_t) \tag{17}$$

$$h_t = o_t * \tanh(c_t) \tag{18}$$

Synthesis


Input Gate
Output Gate
Cell State from
timestamp = t - 1
Cell State to
timestamp = t + 1
Forget Gate
σ
σ
tanh
σ
tanh
Hidden state input
(from
timestamp = t - 1)
Hidden state
output for
timestamp = t + 1
Input Data;
Timestamp = t

CNN-LSTM Model

Scientific Project

CNN in our case

- CNNs designed for image processing.
- Application to stock prediction by recognizing local patterns.
- Conceptualizing time-series data as a 1D image.

1D CNN Architecture

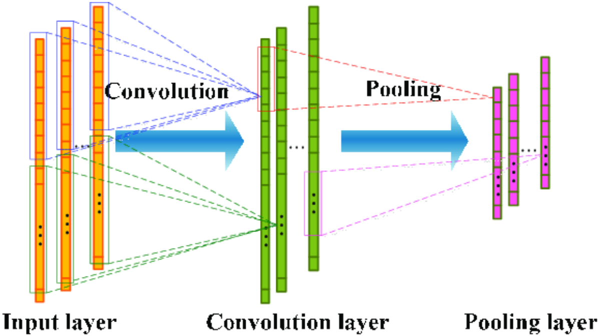


Figure: CNN-LSTM architecture

1D CNN Architecture

- **Input:** X – 1D time-series data.
- **Convolution:**
$$Y = f(X * W + b)$$
- **Activation:**
$$Z = \text{ReLU}(Y)$$
- **Pooling:**
$$P = \text{MaxPooling}(Z)$$

CNN-LSTM Fusion

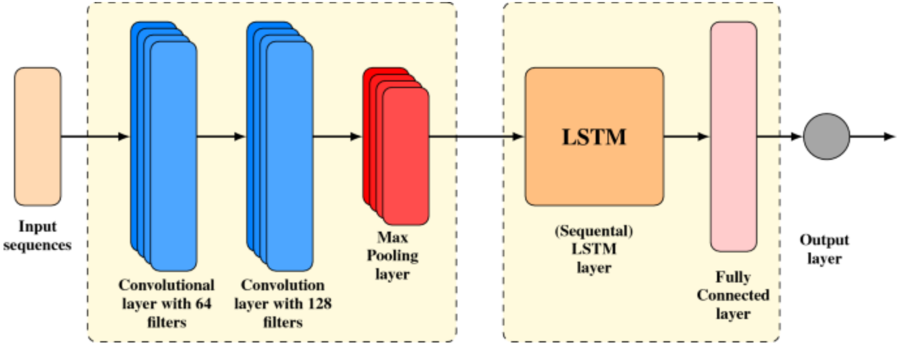


Figure: CNN-LSTM architecture

CNN-LSTM Fusion

- **CNN Output:**

  $$O_{\text{CNN}}$$

  Captures spatial features.
- **LSTM Input:**

  $$I_{\text{LSTM}} = [O_{\text{CNN}}, X]$$

  Concatenation of CNN output and original data.
- **LSTM Output:**

  $$O_{\text{LSTM}}$$

  Handles temporal dependencies.

Application

Scientific Project

Data Overview : Apple Stock

Date	Open	High	Low	Close	Volume
2024-01-08 00:00:00-05:00	182.089996	185.600006	181.500000	185.559998	59144500
2024-01-09 00:00:00-05:00	183.919998	185.149994	182.729996	185.139999	42841800
2024-01-10 00:00:00-05:00	184.350006	186.399994	183.919998	186.190002	46792900
2024-01-11 00:00:00-05:00	186.539993	187.050003	183.619995	185.589996	49128400
2024-01-12 00:00:00-05:00	186.059998	186.740005	185.190002	185.919998	36923605

Figure: Apple's Data

Apple Stock Trend

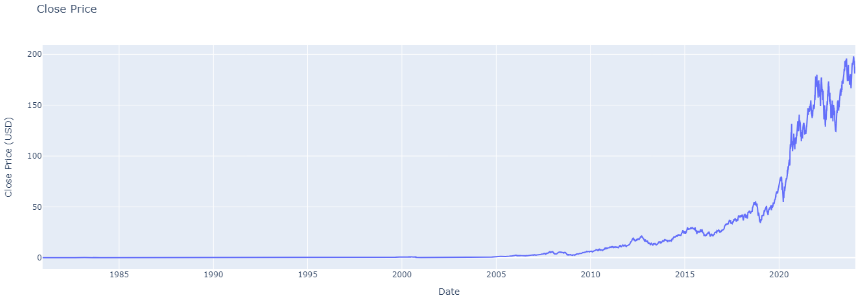


Figure: Apple's Trend

Results

Scientific Project

ARIMA result

Modèle ARIMA - Prédiction vs Réel pour AAPL

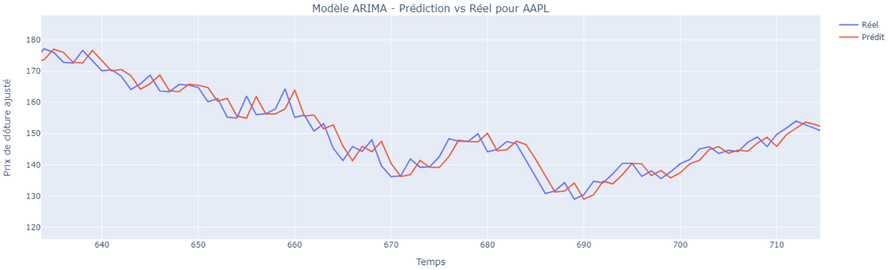
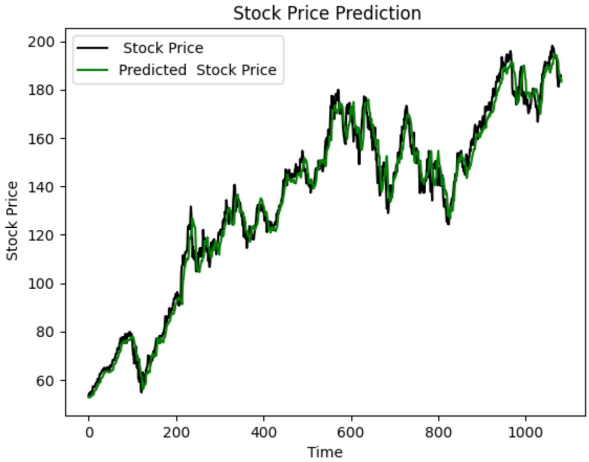


Figure: Apple's Data

LSTM result

Figure: Apple's Data

CNN-LSTM result



Evaluation

Model	MSE	MAE	RMSE
CNN-LSTM	29.214037	4.24860	5.4050
LSTM	7.77e+04	6.9574	8.8133
ARIMA	54.2668	5.5958	7.3666

Table: Different metrics of the models

Comparing the models

Criteria :

- For MSE and MAE, lower values are better.
- For RMSE, lower values are also better.

Comparison : Based on these criteria, the CNN-LSTM model appears to be the most performant.

Conclusion

References

- H.Q.Thang, *Vietnam Stock Index Trend Prediction using Gaussian Process Regression and Autoregressive Moving Average Model*, Research and Development on Information and Communication Technology, HUST, 2018.
- Kim T, Kim HY, *Forecasting stock prices with a feature fusion LSTM-CNN model using different representations of the same data*, PLoS ONE 14(2): e0212320, 2019.
- Hao Y, Gao Q, *Predicting the Trend of Stock Market Index Using the Hybrid Neural Network Based on Multiple Time Scale Machine Learning*, MDPI Appl. Sci. 2020, 10(11), 3961.
- CS231n, *Convolutional Neural Networks (CNNs / ConvNets)*, `https://cs231n.github.io/convolutional-networks/`.
- Aston Zhang, Zachary C. Lipton, *Dive into Deep Learning*.

Thank you for your attention

Y.Hiraji, B.Hajji, Y.Lourarhi, A.Mansouri, R.Nazih

Dr. AHIDAR Adil
Ecole Centrale Casablanca

January 19, 2024