Streaming application for predictions of markets stocks

Quentin Lesbre, Lounes Guergous, Xavier Brouty

Lecturers : Mariam BARRY, Jérémie SUBLIME, Sathiya KUMAR, Maurras TOGBE

16 January 2023

Table of Contents

Table of Contents

Introduction

- The objective of this work is to use online regression with River to predict markets stocks returns.
- We decided to study the following stocks : Meta, BNP Paribas, Alibaba, Gazprom, Saudi Arabian Oil Company and Fiat Chrysler.
- We used algorithms from the regression and classification literature to predict the returns of the markets stocks.
- Bonus : we implemented the indicator from : A statistical test of market efficiency based on information theory, X.Brouty, M.Garcin, to see how it works with streaming data.
  We also developed a really simple portfolio construction strategy to see the performance of the prediction in the markets.

Table of Contents

Batch Classification

- We started with the batch classification with two algorithms : the gradient boosting and the KNN Classifier. We did some features engineering to compute some interesting indicators to give us some insights about the sime series. We have the following results:

	META	BABA	BNP	2222	FCAU	GSPC
Lag 2, Volume et sans aroon	0.309192	0.232774	0.338348	0.340166	0.257618	0.292804
Lag 2, Volume et avec aroon	0.341219	0.301658	0.346721	0.390895	0.308302	0.329753
Lag 5, Volume et sans aroon	0.334651	0.344678	0.402463	0.352227	0.257618	0.269188
Lag 5, Volume et avec aroon	0.323835	0.352063	0.328499	0.437283	0.351601	0.366214

(a) Balanced accuracy for KNN Classifier

	META	BABA	BNP	2222	FCAU	GSPC
Lag 2, Volume et sans aroon	0.329739	0.286462	0.306176	0.349026	0.340563	0.333431
Lag 2, Volume et avec aroon	0.295003	0.310412	0.390214	0.328175	0.346004	0.393368
Lag 5, Volume et sans aroon	0.330687	0.324900	0.334955	0.345992	0.330846	0.388520
Lag 5, Volume et avec aroon	0.322222	0.329071	0.336114	0.330673	0.300295	0.410742

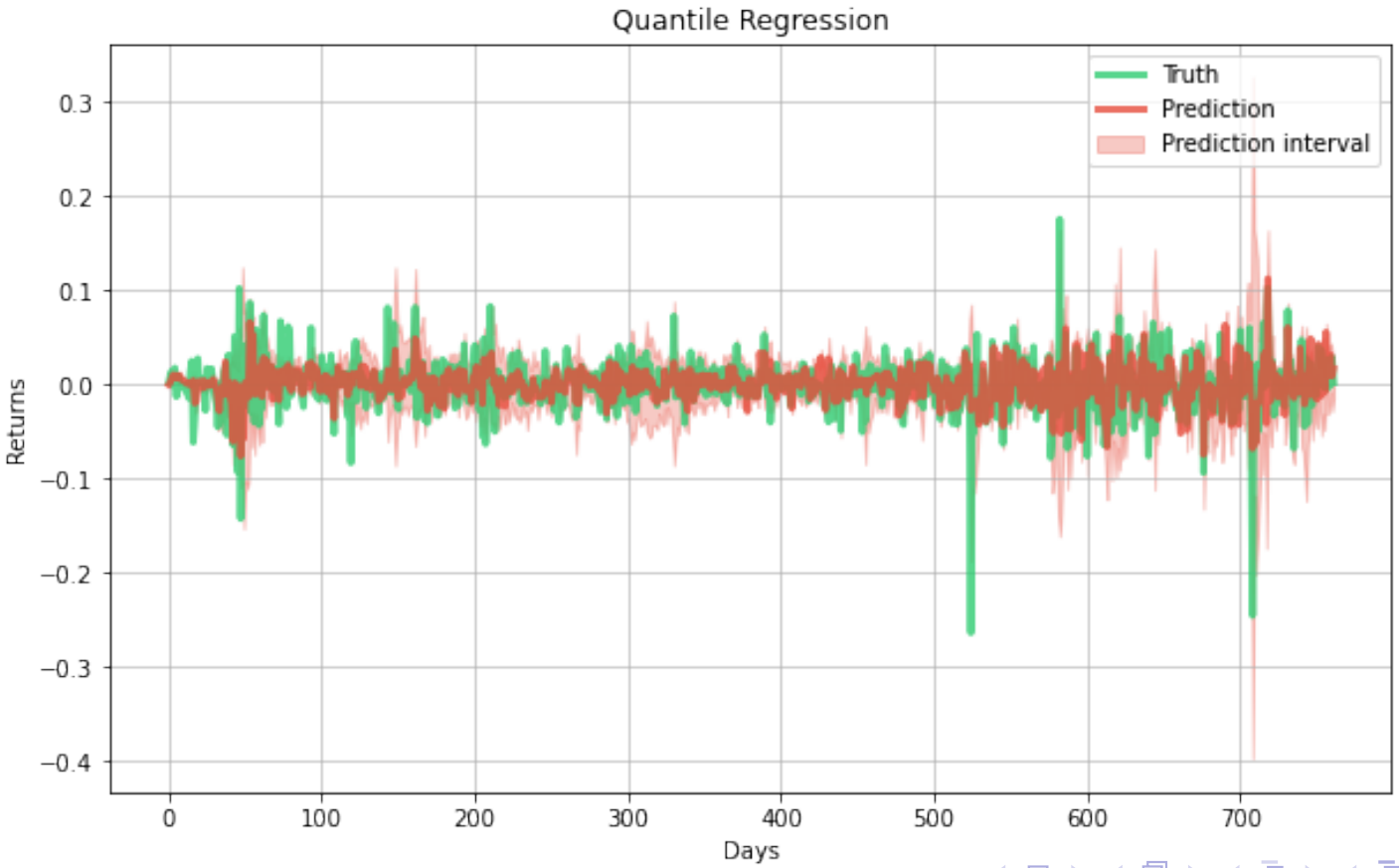
(b) Balanced accuracy for Gradient Boosting

- We can see that some results are quite good compared to randomness depending on the number of lags we take in the features and if we have the aroon indicator or not.

Table of Contents

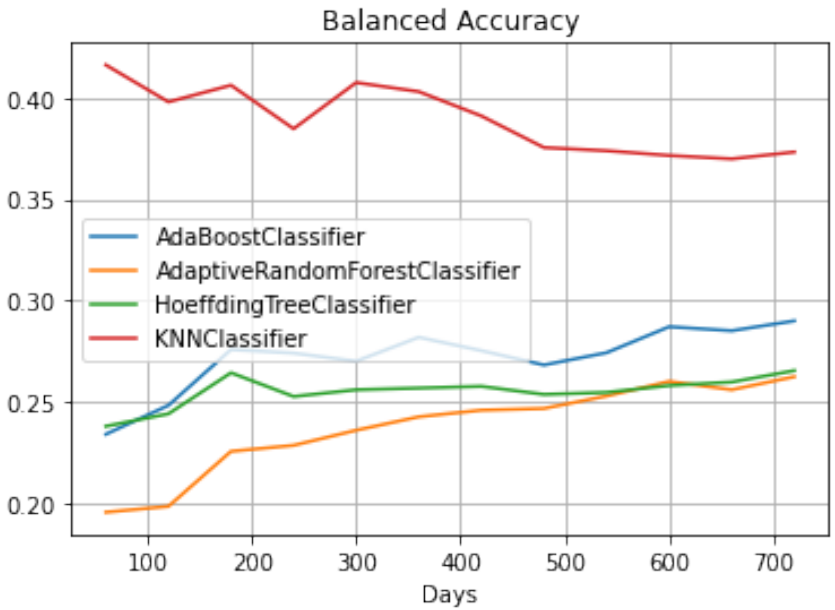
Regression with River

- We started doing regression to predict the future returns of the stocks, because predicting a single value is not really accurate and can be restrictive while taking a decision about the future returns. We did a quantile regression to have an idea of the distribution of the future returns instead of a single value. The interval can also be interesting for risk management where ones want to predict the potential value of a negative returns for a portfolio of stocks.



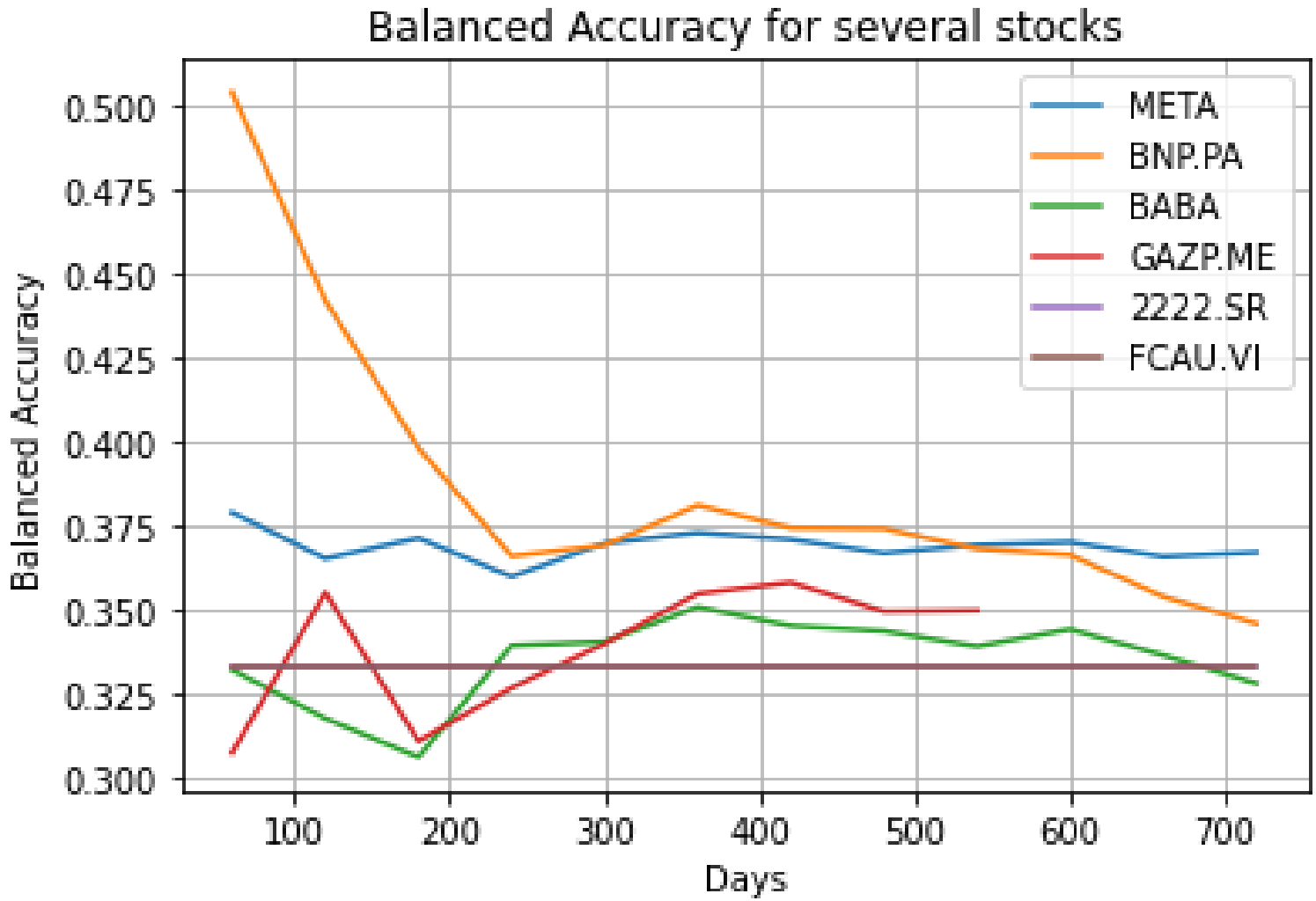
Classification with River

- We then did classification with the River library. We worked with three classes (we note P_t the close price of day t):
  - 0 if $\frac{P_{t+1}-P_t}{P_t} \leq \alpha_1$
  - 1 if $\alpha_1 < \frac{P_{t+1}-P_t}{P_t} < \alpha_2$
  - 2 if $\frac{P_{t+1}-P_t}{P_t} \geq \alpha_2$
- We implemented the function *evaluate binary* to create the streaming data and apply the different models (*KNNClassifier*, *HoeffdingTreeClassifier*, *AdaBoostClassifier*, *AdaptiveRandomForestClassifier*) (see the github repo for more precisions of the code). We also added the features coming from the prediction with quantile regression above. We have the following figure (with META stock):



Classification with River

- We see that the KNN Classifier has the better balanced accuracy so we are going to use this model with the other stocks.



Bonus

- We started by implementing the indicator of market information coming from [1][1]. The formula is the following :

$$I^{L+1} = H_{\star}^{L+1} - H^{L+1} \tag{1}$$

with :

$$H_{\star}^{L+1} = -\sum_{i=1}^{2^L} p_i^L \log_2\left(\frac{p_i^L}{2}\right) \tag{2}$$

$$H^{L+1} = -\sum_{i=1}^{2^L} (p_i^L \pi_i^L \log_2(p_i^L \pi_i^L) + p_i^L (1-\pi_i^L) \log_2(p_i^L (1-\pi_i^L))) \tag{3}$$

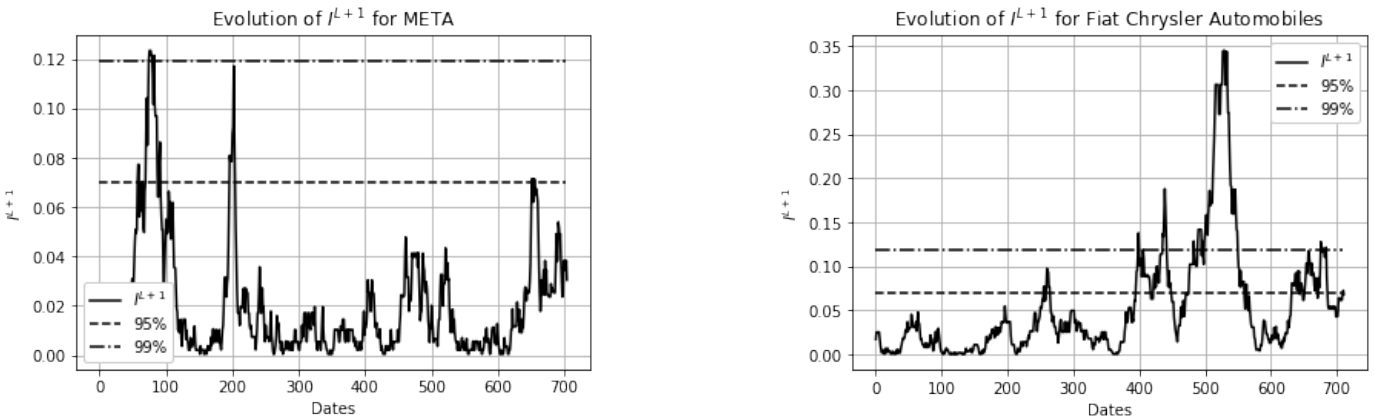
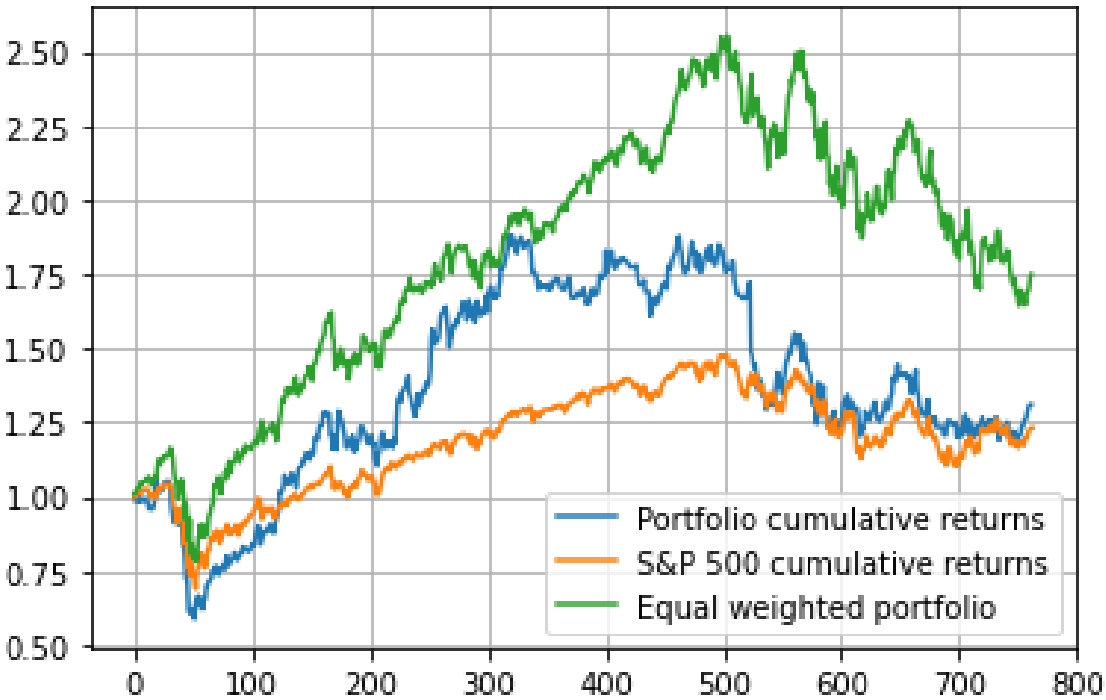


Figure: Information for META and Fiat

[1] [1] https://arxiv.org/abs/2208.11976v1, X.Brouty, M.Garcin

Construction of a portfolio

For the moment, we did the classification and the regression for the stock META, but it is important to construct a portfolio with several stocks in order to improve the performance of the strategy by reducing the risks for a single stock. We are then doing a prediction at each end of the day for the n stocks of the universe of investment and we buy the stocks with a prediction of 2. We will then run a backtest to see if this really simple strategy is working.

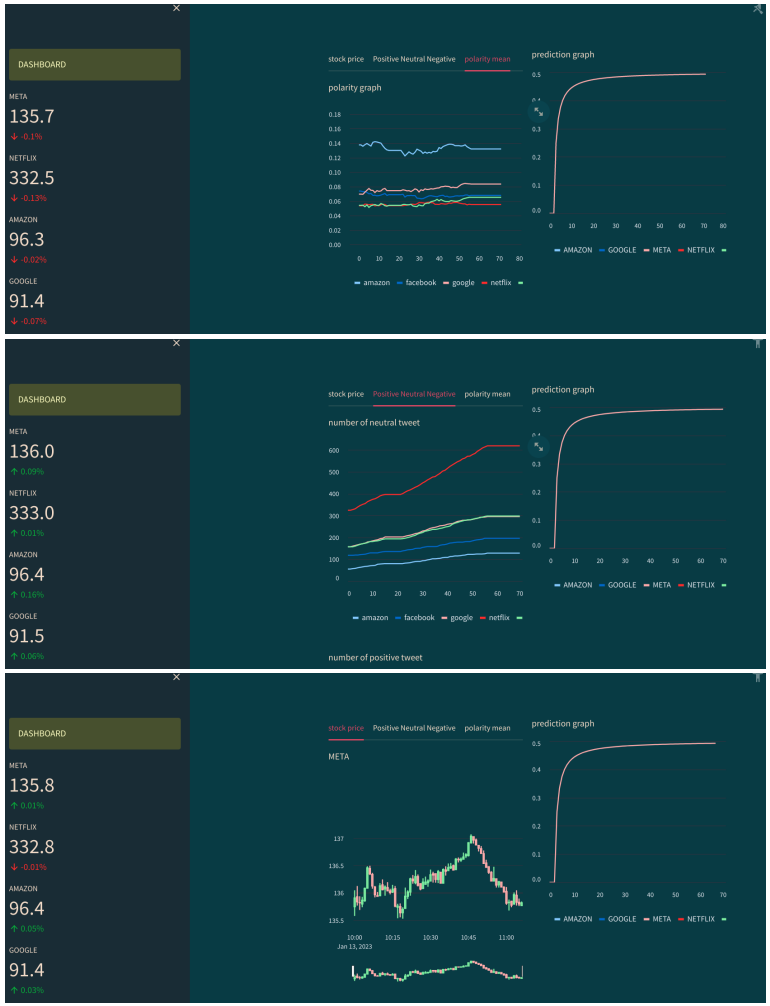


(a) Cumulative Returns for the S&P 500, the portfolio and an equal weighted portfolio of the stocks in the universe of investment

Table of Contents

Streaming application

- For the streaming application with Kafka, we have the following graphs :



Streaming application

- The above graphs are created from data coming in streaming from Kafka and the yahoo finance API. Every minute, a new financial data corresponding to the price of the stocks is given to the algorithm. We can also observe the tweets coming at the same time with the keyword corresponding to the name of the stock, in order to have a sentiment analysis of the traders on this stock. We combine both approaches and it gives us a prediction for the price of the next minute.

Questions ?