INSTITUT
POLYTECHNIQUE
DE PARIS

Data Stream Processing

Stock market forecasting application with Kafka

Stream vs Batch learning

Mohamad EL OSMAN
Mohammed JAWHAR

1 Introduction

Introduction

Project Structure

By harnessing Kafka architecture and advanced streaming models, we delivered real-time stock closing price forecasts.The implementation of a live dashboard facilitated dynamic visualization . In addition , our analysis compared batch and streaming models, highlighting their strengths in stock forecasting.

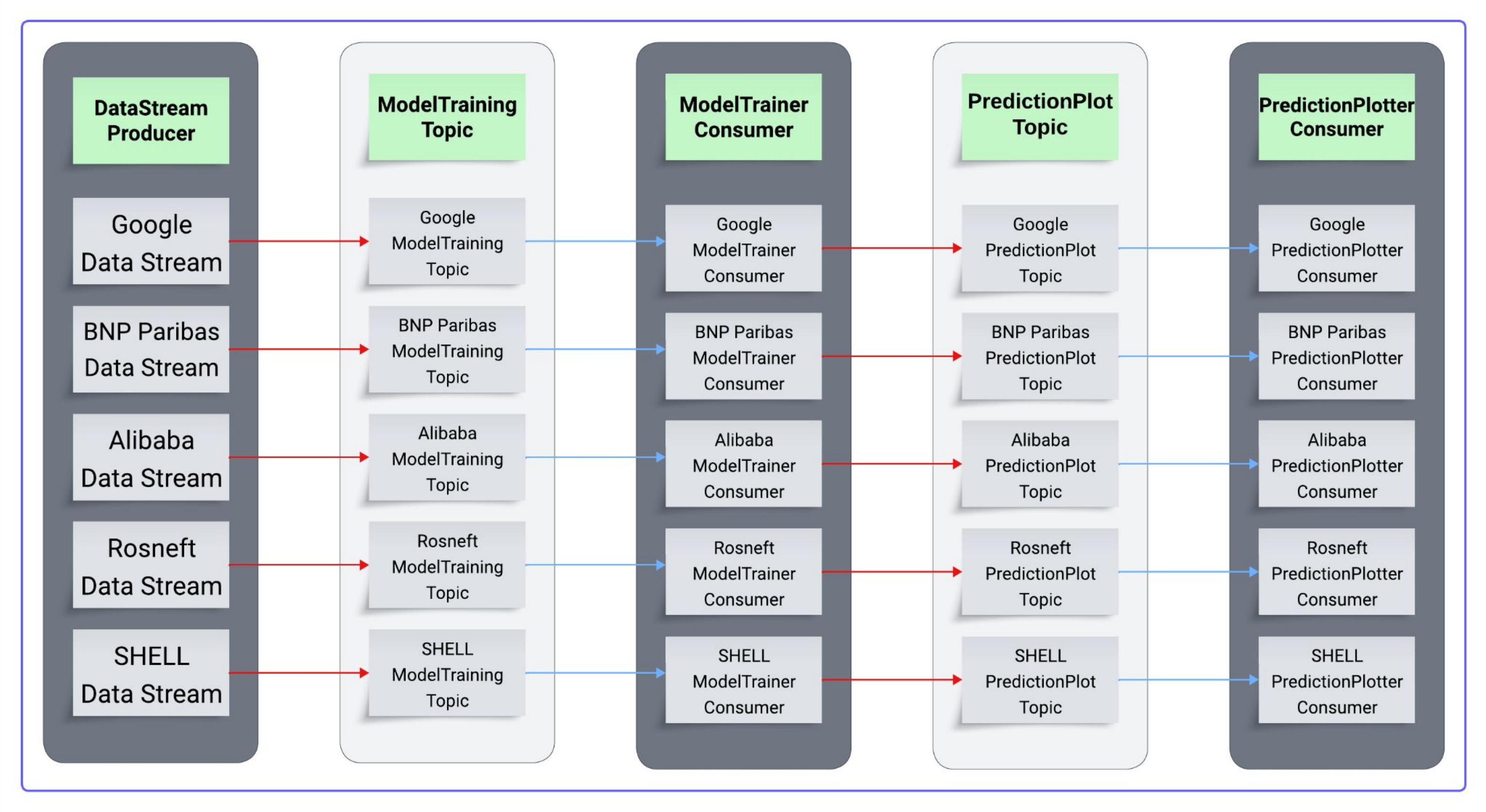
Data Source

Our initiative involved sourcing stock data from the yfinance API, encompassing five prominent companies, each representing a major entity within its respective country.

2 Stream learning

- Kafka Architecture
- Streaming Models
- Metrics Analysis
- Demo

1 - Kafka Architecture



2 – Streaming Models

Linear Regression Model

- Accommodating mini-batch learning, proves efficient for real-time and streaming data processing .
- Computational simplicity facilitates continuous coefficient updates with incoming data, enabling dynamic forecasting .

SARIMAX Model

- Combines various time series models in a unified, online trainable framework .
- Integrates seasonal, non-linear (allowing various regression models), autoregressive, differencing (for integration), moving average, and exogenous inputs .

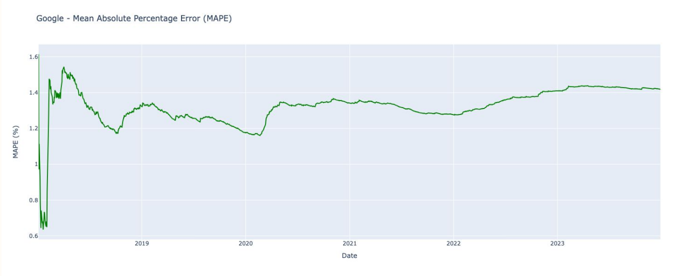
SRP Regression Model

- Specialized models improve stock forecasts via diverse data subsets .
- Versatile learners expand adaptability beyond traditional methods .

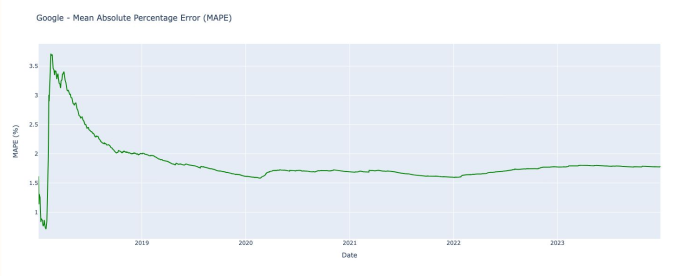
3 – Metrics Analysis

1- Linear Regression Model

2- SARIMAX Model

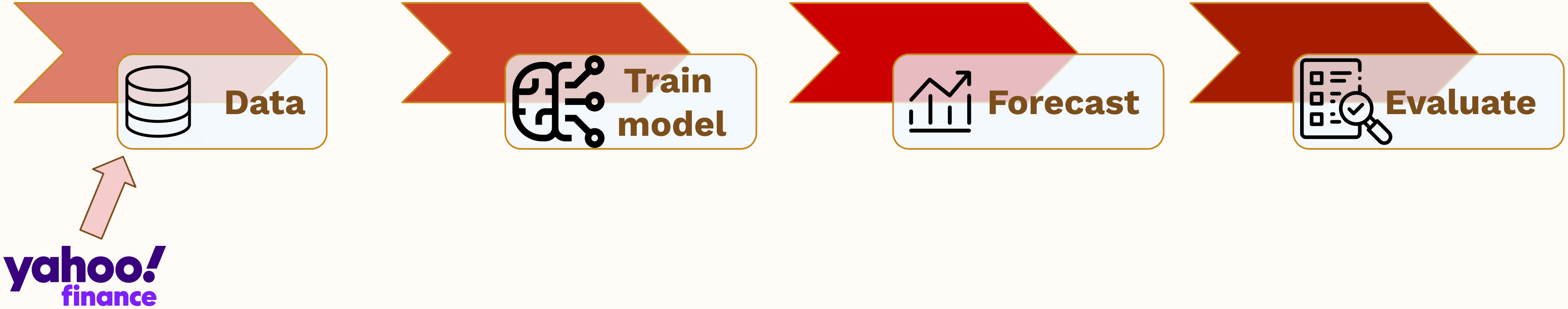


3- SRP Regression Model

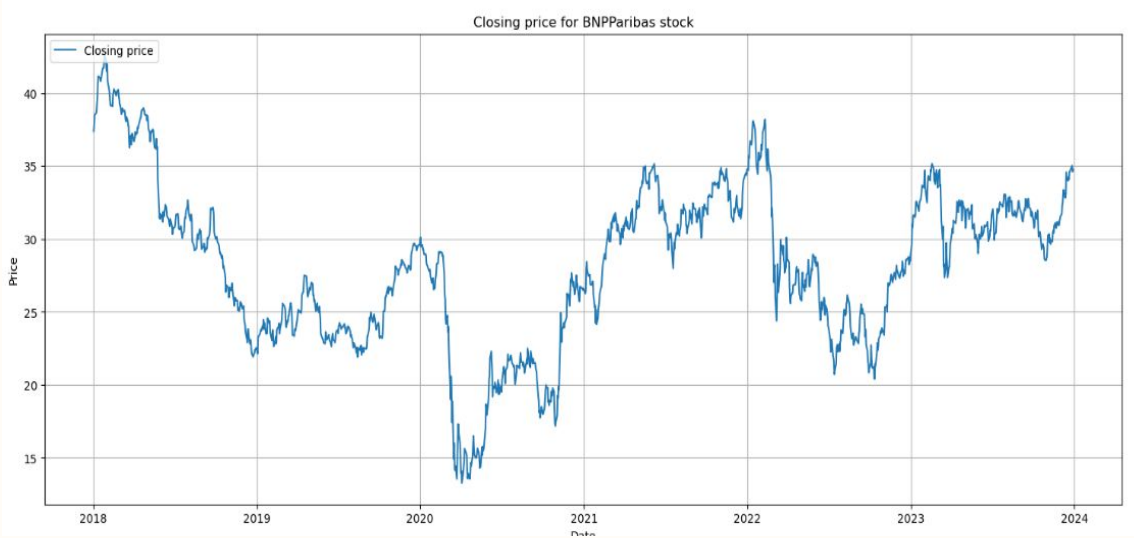


4 – Demo

Batch learning

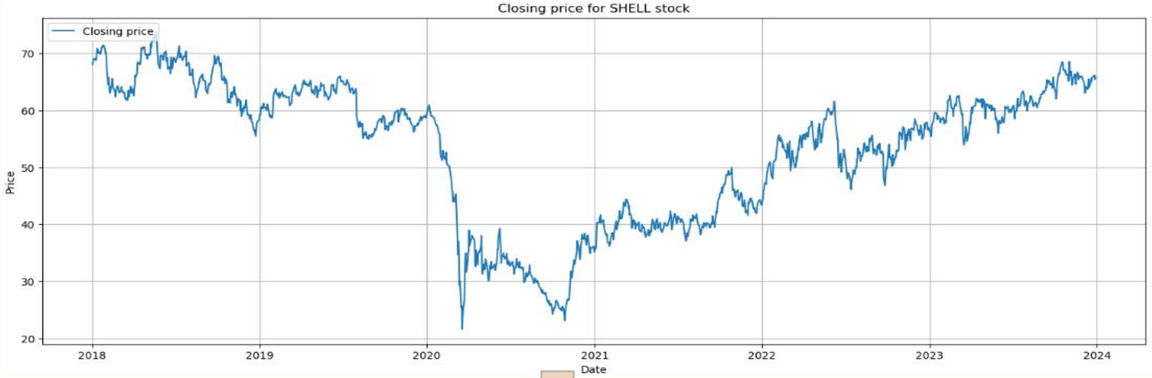


Time series exploration

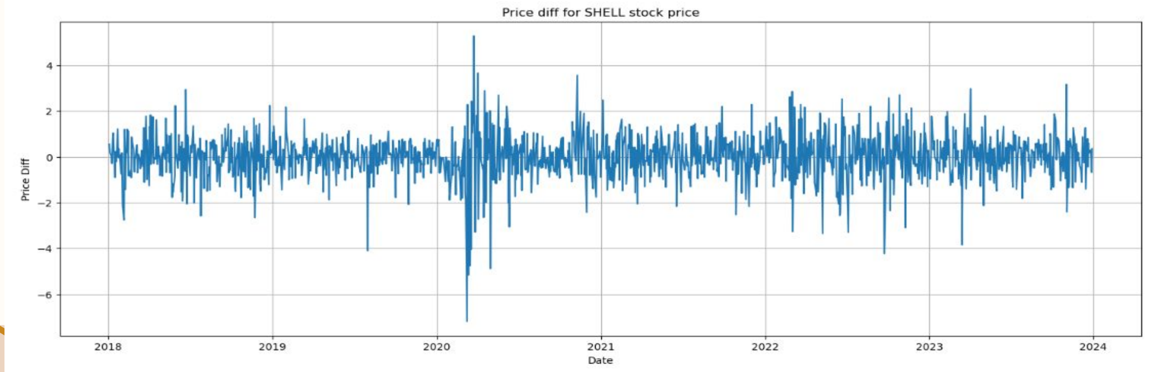


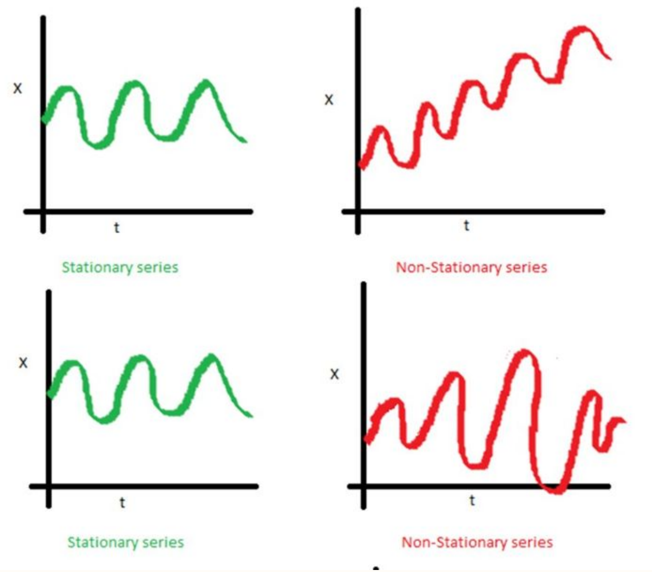
- Stationarity
- ACF and PACF
- Seasonality

Stationarity : Why do we need our stocks to be stationary ?



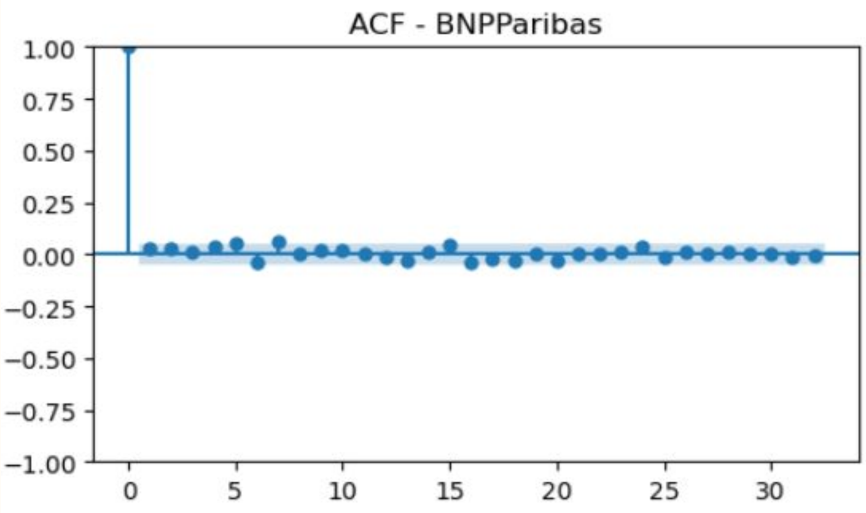
Differencing



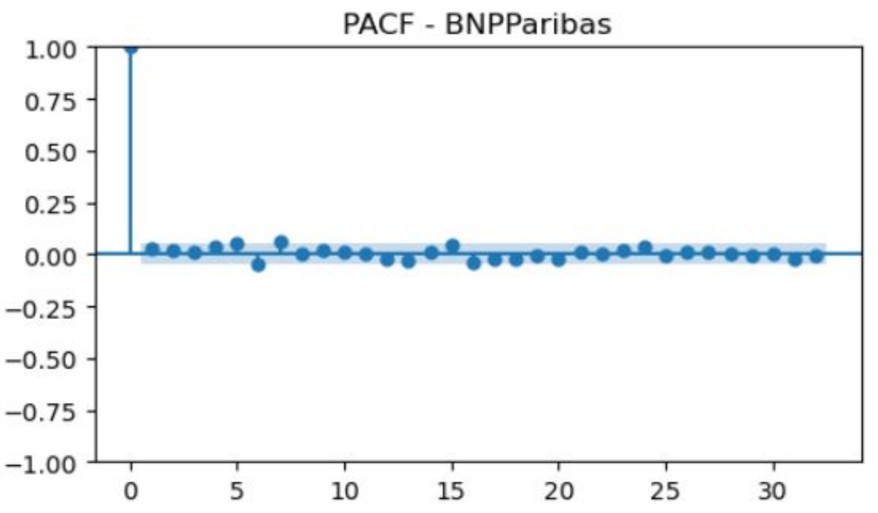


ACF

PACF



ACF - BNPParibas
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
5
10
15
20
25
30



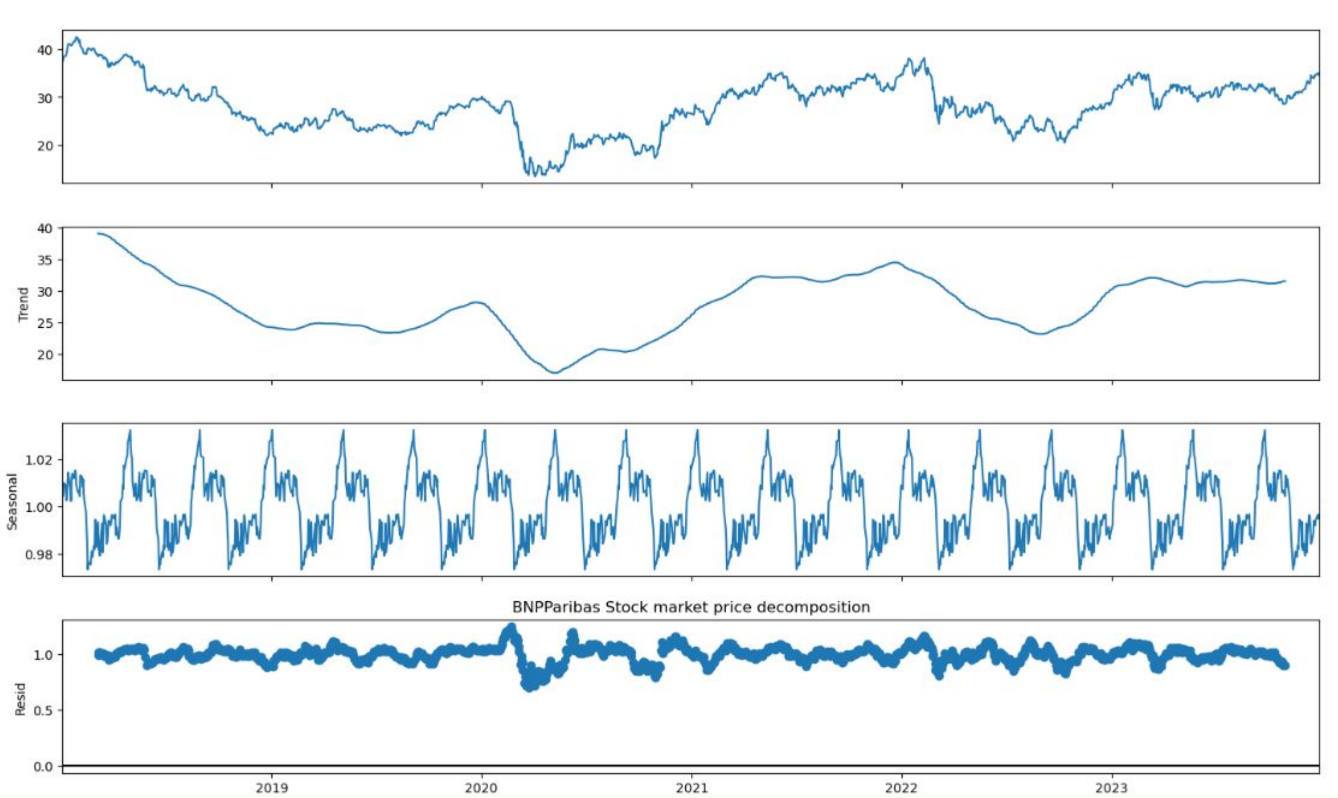
PACF - BNPParibas
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
5
10
15
20
25
30

Seasonality

Multiplicative seasonality :

Y[t] = T[t] * S[t] * e[t]

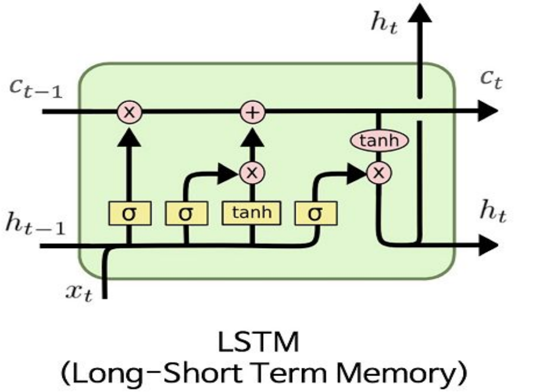
Y[t]: Stock market time-series
T[t]: Trend
S[t]: Seasonality
e[t]: Residual



Batch Modeling

- LSTM
- ARIMA
- PROPHET

LSTM



Company	Model	MAE	MAPE	RMSE	Training_time
Alibaba	LSTM	11.069003	12.263301	15.022452	83.284258
BNPParibas	LSTM	4.489413	13.881668	5.074188	88.704041
Google	LSTM	34.337434	26.914964	38.620671	79.328227
ROSNEFT	LSTM	114.886507	20.139098	132.038403	53.108337
SHELL	LSTM	10.224265	16.334501	10.659568	82.276858

ARIMA

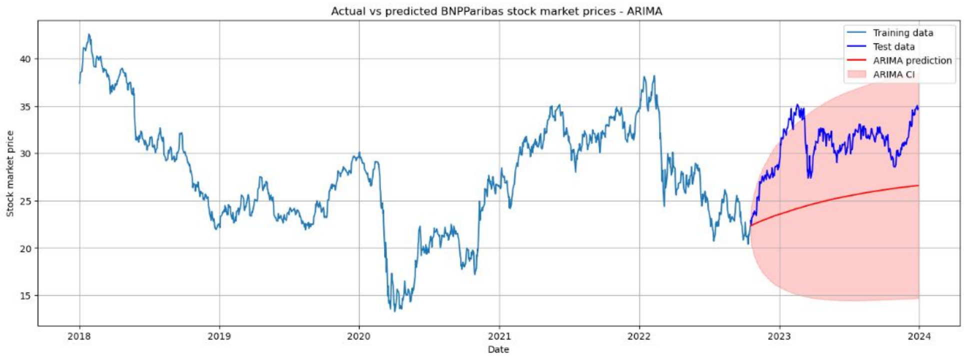
AR
AutoRegressive(p)

I
Integrated

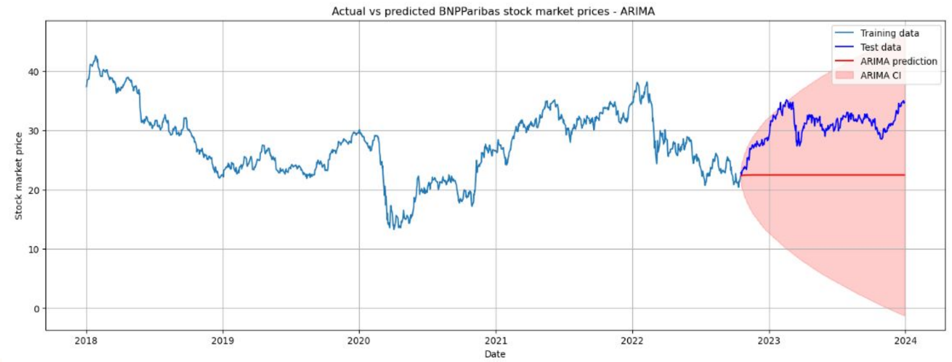
MA
Moving Average(q)

Company	Model	MAE	MAPE	RMSE	Training_time
Alibaba	ARIMA	12.724367	13.606417	15.843760	0.195274
BNPParibas	ARIMA	8.288144	26.364289	8.679870	0.265139
Google	ARIMA	18.919701	15.192441	23.085878	0.157860
ROSNEFT	ARIMA	64.801035	13.513756	82.475474	0.109774
SHELL	ARIMA	9.293301	15.005175	9.982132	0.268873

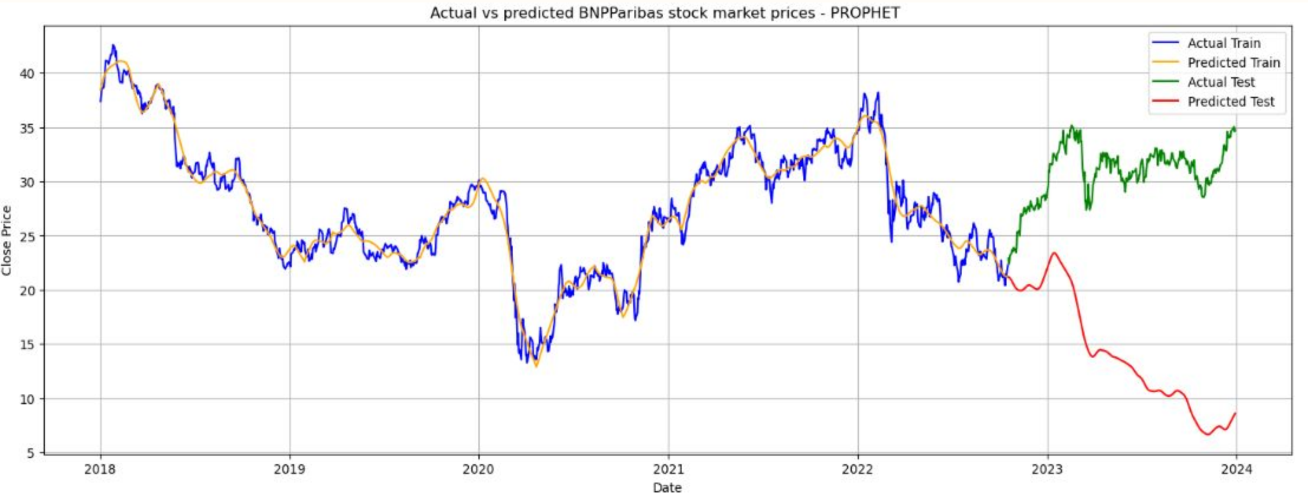
p=1, d=0 , q=1



p=1, d=1 , q=1



PROPHET



Company	Model	MAE	MAPE	RMSE	Training_time
Alibaba	PROPHET	90.734086	104.298664	93.038018	1.484017
BNPParibas	PROPHET	19.773632	63.578232	20.294680	1.061626
Google	PROPHET	50.907372	41.389943	58.202986	0.956702
ROSNEFT	PROPHET	238.393972	48.870004	263.863182	1.019539
SHELL	PROPHET	8.829471	14.000087	11.212083	1.226049

PROPHET

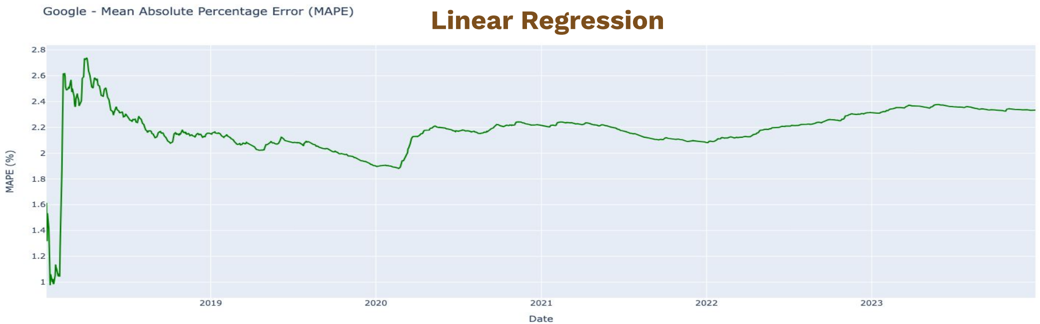
- Prophet is a time series forecasting model developed by Meta's Core Data Science team.
- Designed for forecasting time-series data that exhibit patterns such as seasonality and holidays
- Particularly effective for business applications where datasets may have missing values, outliers, or non-linear trends.

Company	Model	MAE	MAPE	RMSE	Training_time
Alibaba	LSTM	11.069003	12.263301	15.022452	83.284258
BNPParibas	LSTM	4.489413	13.881668	5.074188	88.704041
Google	LSTM	34.337434	26.914964	38.620671	79.328227
ROSNEFT	LSTM	114.886507	20.139098	132.038403	53.108337
SHELL	LSTM	10.224265	16.334501	10.659568	82.276858

Company	Model	MAE	MAPE	RMSE	Training_time
Alibaba	ARIMA	12.724367	13.606417	15.843760	0.195274
BNPParibas	ARIMA	8.288144	26.364289	8.679870	0.265139
Google	ARIMA	18.919701	15.192441	23.085878	0.157860
ROSNEFT	ARIMA	64.801035	13.513756	82.475474	0.109774
SHELL	ARIMA	9.293301	15.005175	9.982132	0.268873

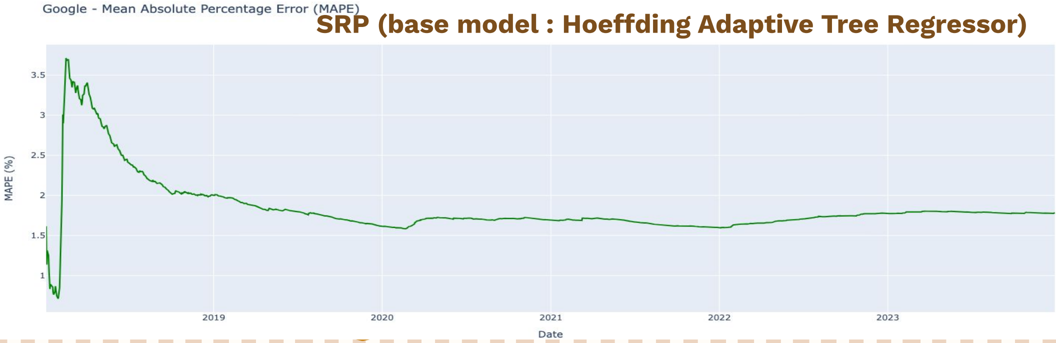
Company	Model	MAE	MAPE	RMSE	Training_time
Alibaba	PROPHET	90.734086	104.298664	93.038018	1.484017
BNPParibas	PROPHET	19.773632	63.578232	20.294680	1.061626
Google	PROPHET	50.907372	41.389943	58.202986	0.956702
ROSNEFT	PROPHET	238.393972	48.870004	263.863182	1.019539
SHELL	PROPHET	8.829471	14.000087	11.212083	1.226049

Linear Regression



Snarimax

SRP (base model : Hoeffding Adaptive Tree Regressor)



Conclusion and perspectives

Batch models

- **Benefit from fitting on all historical data**
- **Struggle in detecting changepoints(abrupt changes)**
- **Non reliable long term predictions**
- **Faster training**

Streaming models

- **Adapt to changes quickly.**
- **Can be sensitive to outliers.**
- **Perform well on a short term**