Dynamic models for multi-variable analysis on stock market behavior

January 11, 2022

Student:
Maarten Peters
12754250

Internal Supervisor:
Dr Chirstian Rodriguez Rivero

External Supervisor:
Dr Julián Antonio Pucheta

Programme:
Data Science

Table of contents

Introduction

Related work

Methodology

Results

Discussion

Conclusion

Questions

Introduction

- COVID-19 pandemic developments
- US & West Texas Intermediate (WTI) market crash
- Economic development

COVID-19 pandemic developments

- Spread of COVID-19
- Uncertainty for daily life/businesses
- Impact on economic development
- Possible government response to mitigate consequences

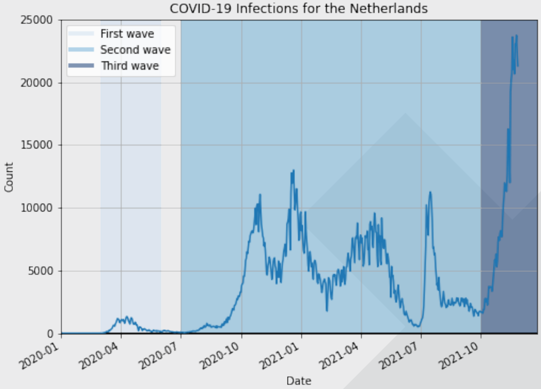


Figure: COVID-19 Infections in the Netherlands, starting from January 1st, 2020

US & WTI market crash

- Indicator of crude oil prices
- Negative drop in first wave
- Proxy of economic development for US
- Consequence: US economy crashed, jobs lost, businesses failing

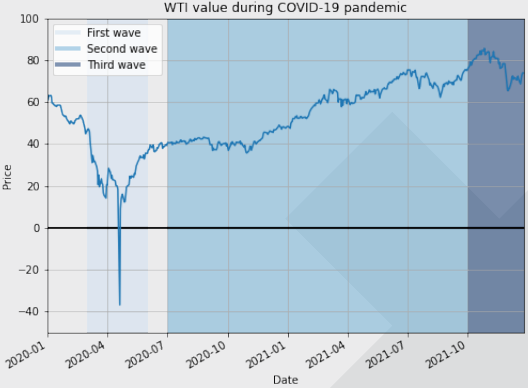


Figure: WTI value during COVID-19 pandemic, starting from January 1st, 2020

Economic development & stock markets

- Stock market indexes (S&P 500 -> US; AEX25 -> NL) as indicator of economic development
- Royal Shell value drops during first wave
- Relationship between COVID-19 and stock market indexes

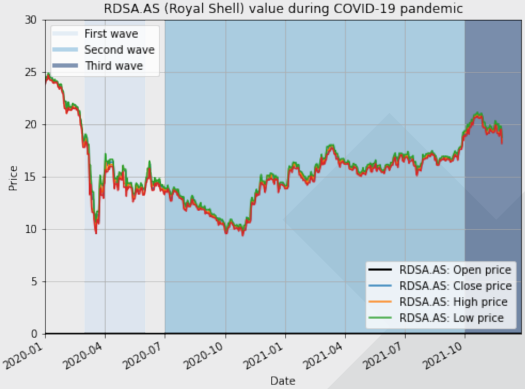


Figure: Royal Shell opening, closing, high and low values during COVID-19 pandemic, starting from January 1st, 2020

Research questions

RQ1 . . . To what extent does the COVID-19 pandemic data influence stock market values?

- Performance of models on short term and long term?
- Model performance on similar data?
- Generalization over different time-periods?
- Which features affect performance to what extent?

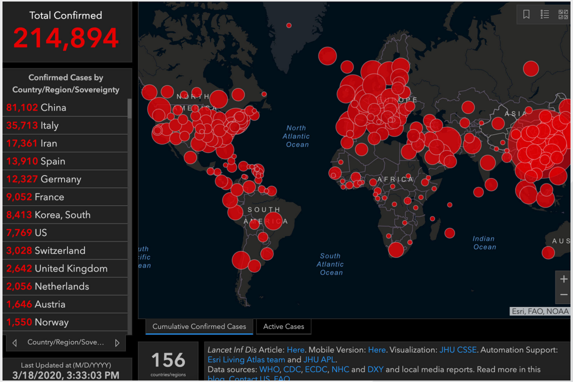
RQ2 . . . To what extent can machine learning techniques aid in prediction?

Related work

- COVID-19 pandemic data collections
- Stock market & environmental data
- Existing research

COVID-19 pandemic data

- **Features:** Infections, deaths, vaccinations & government measures
- Global data by Johns Hopkins University's (JHU) [DDG20] & University of Oxford [Hal+20]
- Local data by the RIVM



Total Confirmed

214,894

Confirmed Cases by Country/Region/Sovereignty

81,102	China
35,713	Italy
17,361	Iran
13,910	Spain
12,327	Germany
9,052	France
8,413	Korea, South
7,769	US
3,028	Switzerland
2,642	United Kingdom
2,056	Netherlands
1,646	Austria
1,550	Norway

Last Updated at (M/D/YYYY)
3/18/2020, 3:33:03 PM

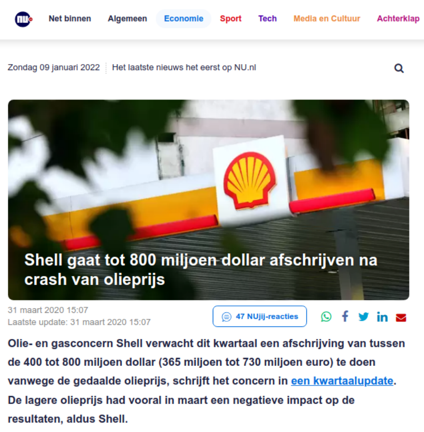
156 countries/regions

Lancet Inf Dis Article: Here. Mobile Version: Here. Visualization: JHU CSSE. Automation Support: Esri Living Atlas team and JHU APL.
Data sources: WHO, CDC, ECDC, NHC and DXY and local media reports. Read more in this blog. Contact US. FAQ.

Figure: JHU COVID-19 real-time dashboard, source: The Philidelphia Inquirer, March 19, 2020

Stock market & environmental data

- **Features:** Open, close, high & low asset prices; traded volume
- Collected from Yahoo Finance API [1]
- Trading and influence by environment:
  - Weather [HS03]
  - News [TK74; DT85; Ver15]



Zondag 09 januari 2022 Het laatste nieuws het eerst op NU.nl

Shell gaat tot 800 miljoen dollar afschrijven na crash van olieprijs

31 maart 2020 15:07
Laatste update: 31 maart 2020 15:07

Olie- en gasconcern Shell verwacht dit kwartaal een afschrijving van tussen de 400 tot 800 miljoen dollar (365 miljoen tot 730 miljoen euro) te doen vanwege de gedaalde olieprijs, schrijft het concern in een kwartaalupdate. De lagere olieprijs had vooral in maart een negatieve impact op de resultaten, aldus Shell.

Figure: Royal Shell amortizes 800 million, source: NU.nl, published March 31st, 2020; taken January 9th, 2021

[1] Python package `yfinance`

Existing research

- Existing models for long-term, needing validation [Chu+20; BD20; Fer20]
- Historic pandemics offer insight, but model years/decades [Ost17; CLV18; JST20]
- Short-term models show promise on regional level, government measures and with machine learning [Zha+20; Deb+20; Car+21]

Methodology

- Data: Information gathering
- Models: Choosing and fitting models
- Evaluation: Model evaluation methods

Methodology: Information gathering

Data for the Netherlands was gathered from ...

- COVID-19 pandemic:
  - **JHU**: Vaccinations
  - **Oxford**: Government measures & stringency
  - **RIVM**: Infections & related deaths
- Environmental data, i.e. weather, from the KNMI, containing:
  - Sunshine hours
  - Precipitation
  - Wind speeds

All feature values were compared between results on our error metric.

Models: Choosing and fitting models

Three models were selected with distinct properties for our data features and compared to a baseline:

- **Linear Regression (LR)**: fits if data is linear and independent
- **Random Forest Regressor (RFR)**: fits if data is non-linear
- **ARIMA**: fits if data is non-linear, but dependent
- **Baseline**: returns the last value of time-series, disregarding all input:

$$\hat{Y}_{T+h|T} = Y_T$$

Models were fitted automatically with feature selection

Model Evaluation

Models were compared on mean absolute percentage error (MAPE) for 50 prediction results:

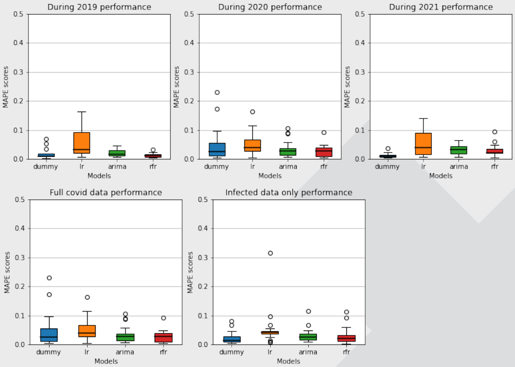
$$\mathrm{MAPE} = \frac{100}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

- If MAPE $>=$ 1.0: Discard the result, as it constitutes a $+100\%$ error
- Baseline error score should be consistent for model comparison

Results: Plots 1

All error scores were compared over different periods, feature sets and daily/weekly frequencies.

- Weekly predictions to inconsistent to use
- All models struggle to outperform baseline



During 2019 performance

During 2020 performance

During 2021 performance

Full covid data performance

Infected data only performance

Figure: Top: Models trained on daily data, split by year, capped at MAPE score <= 0.5
Bottom: Less features favored all models, but no difference from baseline

Results: Plots 2

Possible issues and remarks:

- Sample plots of fitted models showed obvious errors
- Error rate consistently decreased with decreased number of features, regardless of pandemic data or weather data

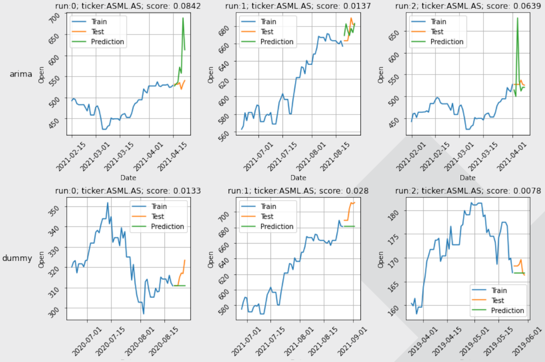


Figure: Sample predictions for ASML opening price, given access to all data features

Results: Significance

Testing for significance of our results, we find that with access to all data features, both state-of-the-art (SOTA) models, ARIMA and RFR, do not significantly ($p < 0.05$) outperform our baseline:

			Baseline comparison	
Featureset	**Model**	**Avg. MAPE**	**t-stat**	**p**
All	**ARIMA**	0.0302	-0.3735	0.7096
	RFR	0.0234	0.6711	0.5037
None	**ARIMA**	0.0315	-1.7365	0.0856
	RFR	0.0242	-0.5310	0.5966

Table: Model MAPE score comparison to baseline, for all dataset features and no additional features

Discussion

Possible reasons for inconclusive results:

- Proper manual model fitting could be necessary, as domain expertise adds to model accuracy
- The efficient market hypothesis (EMH), although debatable, states no individual can outperform the market, given all available information.
- The response of stock markets to the precedent of COVID-19 caused speculation/confusion, which subsided over time.

Conclusion

RQ1 . . . *To what extent does the COVID-19 pandemic data influence stock market values?*
No conclusive evidence that COVID-19 pandemic data influences stock market values.

RQ2 . . . *To what extent can machine learning techniques aid in prediction, given a comparison of baseline models?*
No conclusive evidence that ARIMA/RFR models aid in prediction, models need manual fitting for proper results.

Any questions?

Thank you for your attention!

Additional slides

. . .

Exploratory Data Analysis (EDA) 1

From EDA we learned that ...

- Ordinary least squares regression (OLS) achieves positive R^2 scores ($>$0.7)
- Polynomial models improve on LR, but require manual fitting (score $>$0.9)
- Calendar features over complicate models

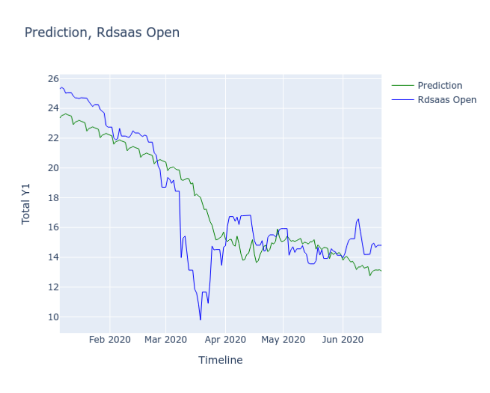


Figure: OLS fit on RDSA opening prices, with weekday features and COVID-19 infections as input

Exploratory Data Analysis (EDA) 2

- Positive relationship heavily dependent on period of time-series
- Correlation scores between RDSA price and COVID-19 infection low, possibly pointing to non-linearity

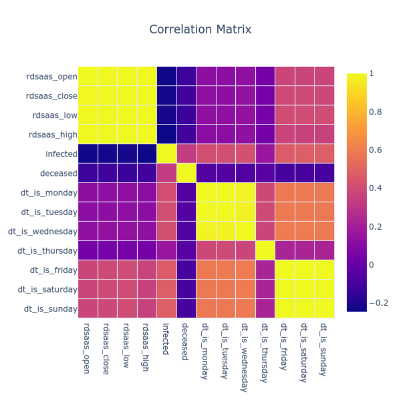


Figure: Heatmap of correlation scores, showing negative correlation between RDSA values and COVID-19 infections and deaths

Preprocessing and Fitting

Before models were fitted, data was sampled and preprocessed in steps:

- Scaling (i.e. between 0 and 1) of all features
- Distribution normalized with Yeo-Johnson transformation [YJ00], a variant of Box-Cox allowing for negative values
- Second scaling

Afterwards, models were trained with recursive feature elimination and five-fold CV on RMSE, with exception for ARIMA

Additional questions 1

- *How well embedded is your research in the existing research?* There is existing research on . . .
  - short term time-series forecasting
  - long term (economic) effects of (COVID-19) pandemic(s)
  - stock market indexes as an indicator of economic development

Additional questions 2

- *Do you deem your study worthy to be published? What is the scientific value of this all?*
  Yes, although the inconclusive results and possible issues from the discussion do devalue it to some degree. It's scientific value comes primarily from . . .
  - (relatively) new topic of time-series forecasting on stock market values with pandemic data
  - insight into automated model fitting

Additional questions 3

- *How did you ensure that the results are reproducible?*
  - All random sampling was seeded consistently
  - Data gathering and preparation was scripted so all data can be generated
  - Data sampling, preprocessing, model training and fitting was scripted, so it could be rerun into infinity
  - Most of the research is documented in the thesis design, EDA and manuscript, along with all code stored in Github.

Additional questions 4

- *Why did you exclude alternative approaches in favour of your current approach?*
  With the pandemic being a global event, most of the effort has gone into investigating the effect of variables and well documented models. A different approach would've been to reduce variables and work towards RNN/ANN/LSTM to improve accuracy, but might reduce generalization. As the latter require large amounts of data and fine-tuning, we felt this was unwarranted given the relative novelty of the COVID-19 pandemic.

Additional questions 5

- *How sound and complete do you consider your data collection criteria?*
  We believe our data collection and preparation is fully complete. There is room for improvement on implementing it on models, specifically on preprocessing and feature selection/engineering, but this should be model driven.

Additional questions 6

- *How did you ensure that the annotation was sound and the ground truth reliable?*
  By putting focus on data gathering and preparation, along with reproducability, we believe our results to be reliable. There is room for improvement on the power of the research, (automatic) (hyper)parameter tuning and quality of the code.

Additional questions 7

- *How is the choice of your current method justified in light of its performance in previous studies?*
  All models are well-documented and have proven their strength and the data also has shown its value in previous studies. The caveat in our study is the automatic model fitting, which did not perform as we hoped, resulting in inconclusive results. Automatic model fitting is to some extent frowned upon, but aides in reproducbility, as rules for fitting are fixed compared to manual intervention.

Additional questions 8

- *How did you address reliability and validity concerns in your research?*
  See slide: Additional questions 3

Additional questions 9

- *How should we interpret the lack of results in your study? Would you say that this was an inherent risk of the set-up you have chosen?*
  Yes, this was an inherent risk of the set-up, but we believe our approach was well considered. This research area was quite novel and had an inherent risk of inconclusive results, given the time, resources and scope of our research. With more time, expertise and an iterative approach on a specific element of our research questions, we might achieve better results.

Additional questions 10

- *How do you explain the differences between your study and other studies?*
  Our study differs as it tries to explain variance in stock market values by utilizing event data, specifically COVID-19 pandemic data. Our EDA hinted this might work and related research, although scarce and novel, did support our hypothesis. The difference to our study is we chose to generalize and automate model fitting instead of engineering a specific model for this specific problem.

Additional questions 11

- *What would you do differently when you could redo your study? Which limitations could have been avoided?*
  - Focus on reprocubility as a last step
  - Base our study primarily on a single related paper/dataset and expand the problem space in smaller increments

Additional questions 12

- *Why can your study be considered a contribution to the field?*
  - It serves as an insight that modelling time-series on stock market values requires extensive domain knowledge and any gained results in that field is a product of the latter, and not necessarily the model or the data.
  - It also demonstrates that coincidental correlation does not mean causation or even a relationship in general.