Wie man eine Repräsentation unsere Planeten erstellt - und von ihr lernt

Prognose von Börsenverhalten basierend auf Ereignissen

Idee

→ Modellierung des Weltgeschehens durch Nachrichten

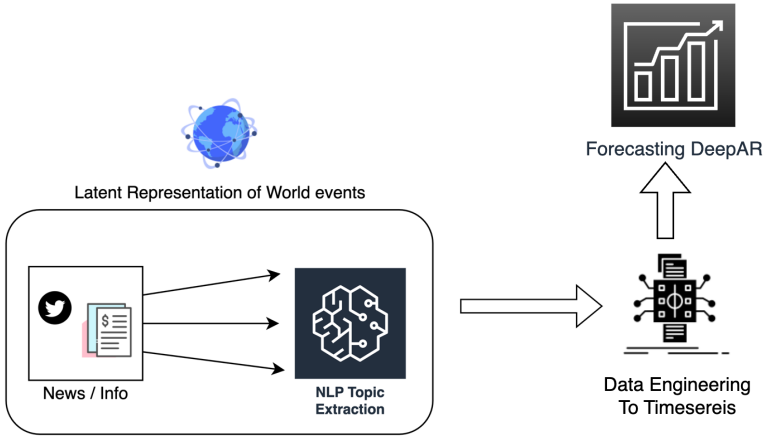
→ Finden von kausalen Zusammenhängen zwischen Nachrichten(typen) und Verhalten von Aktienmarkt

Erster Versuch

- Verwendung von Zeitungsartikeln für wichtigste Ereignisse
- Mit NLP Überschriften in Themen transformieren
- Beziehungen zwischen Themen und Bewegungen am Aktienmarkt finden
- Zukünftige Bewegungen auf Grund von aktuellen Ereignissen vorhersagen

Herausforderungen:

- Nicht genügend Daten
- Nicht konsistent, da viele verschiedene Quellen
- Weltweite Nachrichten in vielen verschiedenen Sprachen

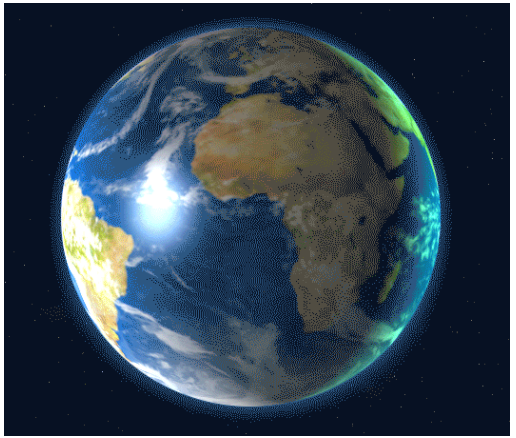


01

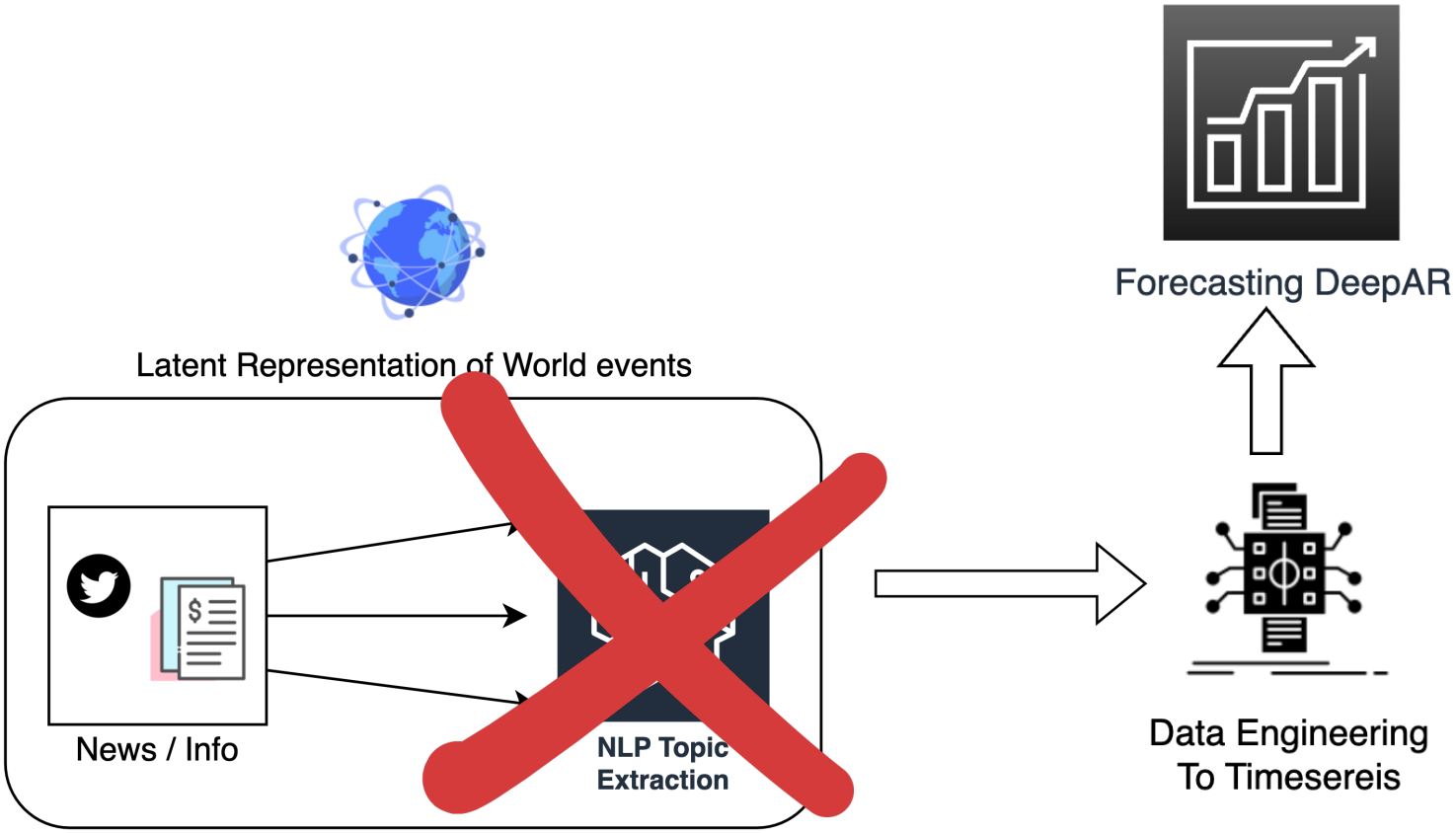
Data Collection (Events)

GDELT Project

"Watching our world unfold"



- Überwacht Nachrichten in diversen Medien in allen Ländern der Welt
- Übersetzt ins Englische
- 1979 - heute
- Jede 15 min aktualisiert
- Datenbank basiert auf Events, nicht Nachrichten (aggregiert)
- Kostenlos verfügbar


Forecasting DeepAR
Latent Representation of World events
News / Info
NLP Topic
Extraction
Data Engineering
To Timesereis

GDELT

Daten in dieser Größe sind nicht leicht lesbar für Menschen.

- 2015 - heute in Google Big Query verfügbar
- 2013 - 2015 downloadbar durch web-crawling
- 1979 - 2013 in sehr großer .txt datei (tab-spaced)

- Die einzelnen Datenpunkte sind in CAMEO verschlüsselt → **C**onflict **a**nd **M**ediation **E**vent **O**bservations (CAMEO)
  - Benötigt Verständnis des CAMEO Konzepts und der Dokumentation
- Wie findet man die wichtigsten Events?
  - Alle Daten wären zu viel (2.5 TB pro Jahr)

GDELT Ergebnis

- Datensatz mit den wichtigsten (max Num Articles) Events der letzten 43 Jahre
  - Spalten: date, event code, average tone, num_articles
- Weltweit

→**Eine große, konsistente und authentische Datengrundlage für die Vorhersage**

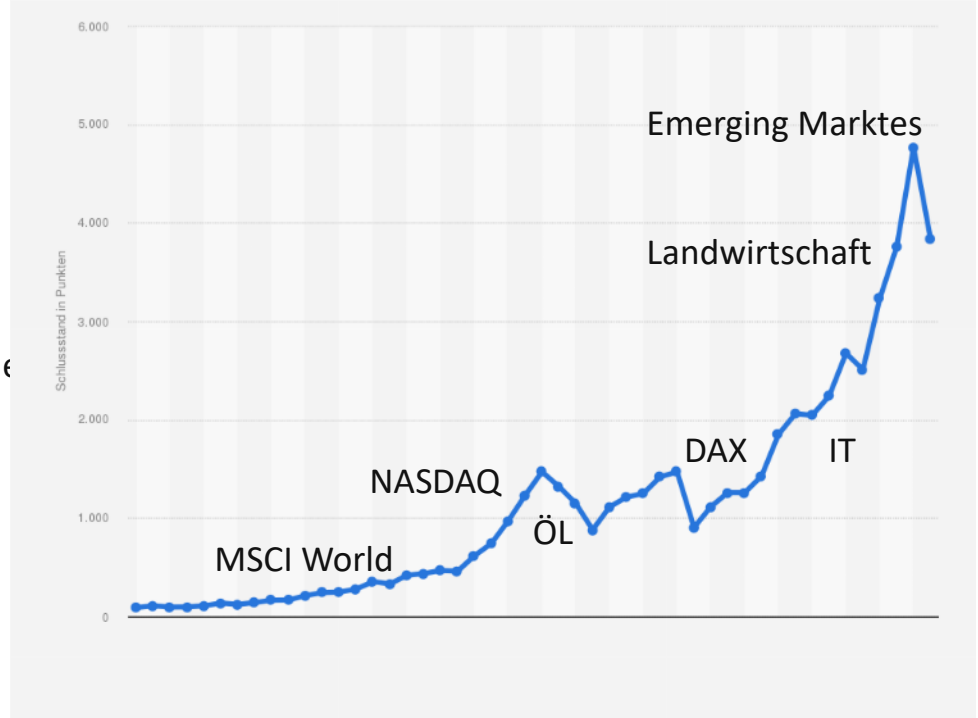
Zeile	EventCode	Actor1Code	Actor1Name	Actor1CountryCode	Actor2Code	Actor2Name	Actor2CountryCode	NumArticles	DATEADDED	AvgTone
1	043	ATG	ANTIGUA AND BARBUDA	ATG	LCA	ST LUCIA	LCA	2	20230606184500	3.030303030303…
2	042	LCA	ST LUCIA	LCA	CRI	COSTA RICA	CRI	2	20230606184500	3.030303030303…
3	042	LCA	ST LUCIA	LCA	URY	URUGUAY	URY	2	20230606184500	3.030303030303…

02

Data Collection (Finanzdaten)

ETF/ Indizes

- Welche ETFs gibt es?
- Welche ETFs waren relevant in Vergangenhe
- Welche Indizes sind aussagekräftig?



Weitere Faktoren

Welche externen Faktoren könnten den Markt beeinflussen?

Leitzins

Inflation

Geldmenge

Arbeitslosigkeit

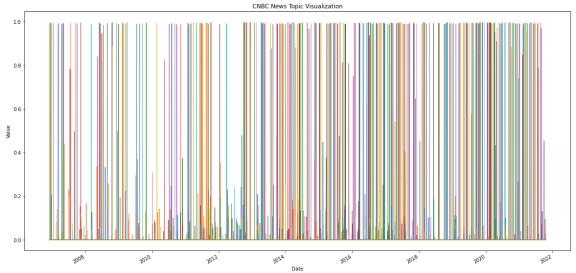
Bruttoinlandsprodukt

03

Data Engineering

Datentransformation

CNBC news



	title	url	published_at
0	Santoli's Wednesday market notes: Could Septem...	https://www.cnbc.com/2021/09/29/santolis-wedne...	2021-09-29T17:09:39+0000
1	My take on the early Brexit winners and losers	https://www.cnbc.com/2016/06/24/ian-bremmers-t...	2016-06-24T13:50:48-0400
2	Europe's recovery depends on Renzi's...	https://www.cnbc.com/2014/03/25/europes-recove...	2014-03-25T13:29:45-0400

NLP Vorverarbeitung

\+

Topic clustering LDA

	published_at	topic_class
0	2021-09-29T17:09:39+0000	[(449, 0.09099977), (561, 0.091001034), (753, ...
1	2016-06-24T13:50:48-0400	[(488, 0.046972986), (562, 0.92542994), (658, ...
2	2014-03-25T13:29:45-0400	[(530, 0.9982684)]
3	2009-04-22T19:49:03+0000	[(161, 0.03850209), (164, 0.09252543), (196, 0...
4	2018-04-14T14:59:04+0000	NaN

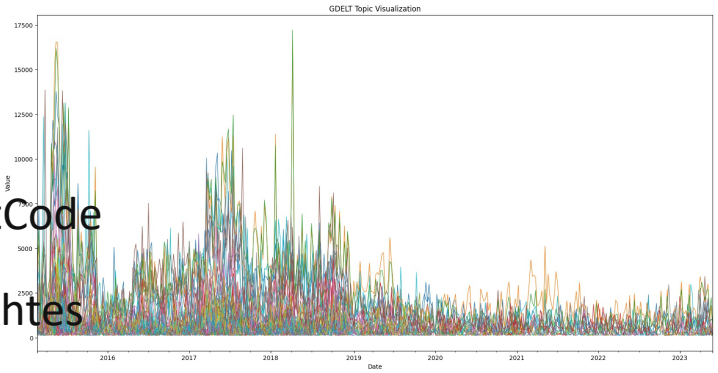
Gruppierung nach Topic & Woche und anschließend Pivotierung in gewünschtes Format

Datentransformation

GDELT

	DATEADDED	EventCode	NumArticles	Goldstein_Scale	AvgTone
0	19790101	51	4	3.4	NaN
1	19790101	10	4	0.0	NaN
2	19790101	46	3	7.0	NaN
3	19790101	46	3	7.0	NaN
4	19790101	51	3	3.4	NaN
...	...	...	...	...	...
9995	19801231	190	4	-10.0	NaN
9996	19801231	841	3	7.0	NaN
9997	19801231	20	3	3.0	NaN
9998	19801231	80	3	5.0	NaN
9999	19801231	195	1	-10.0	NaN

Gruppierung nach EventCode und anschließend Pivotierung in gewünschtes Format



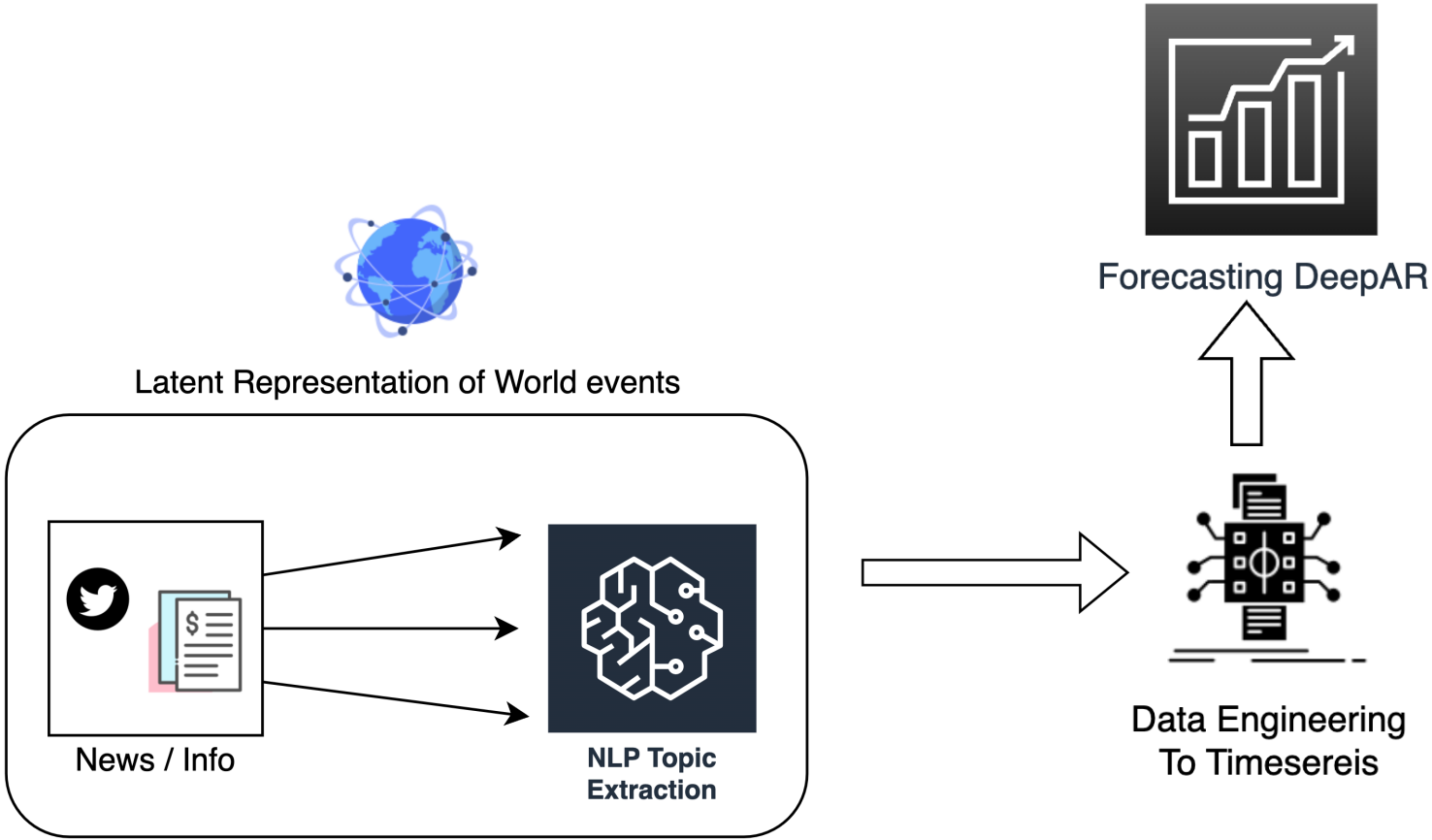
Vorgehen

- Exploratives Arbeiten in Jupyter Notebooks
- Später Umwandlung des Notebook codes in command line kompatible Python Skripte
  - Iteratives Verarbeiten der Skripte aller Experimente

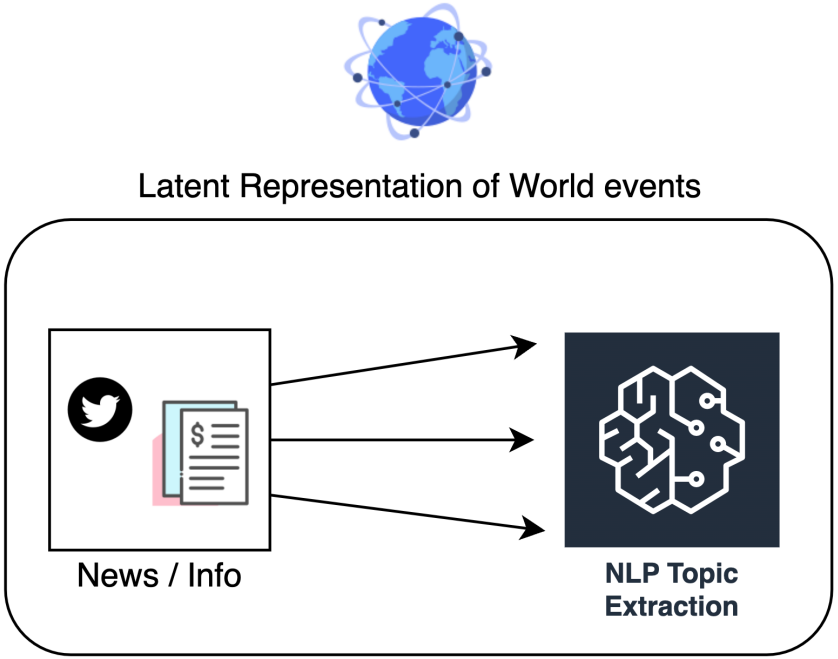

```
py .\timeseries_engineering.py --freq W -i ..modelling/output_data/
```

04

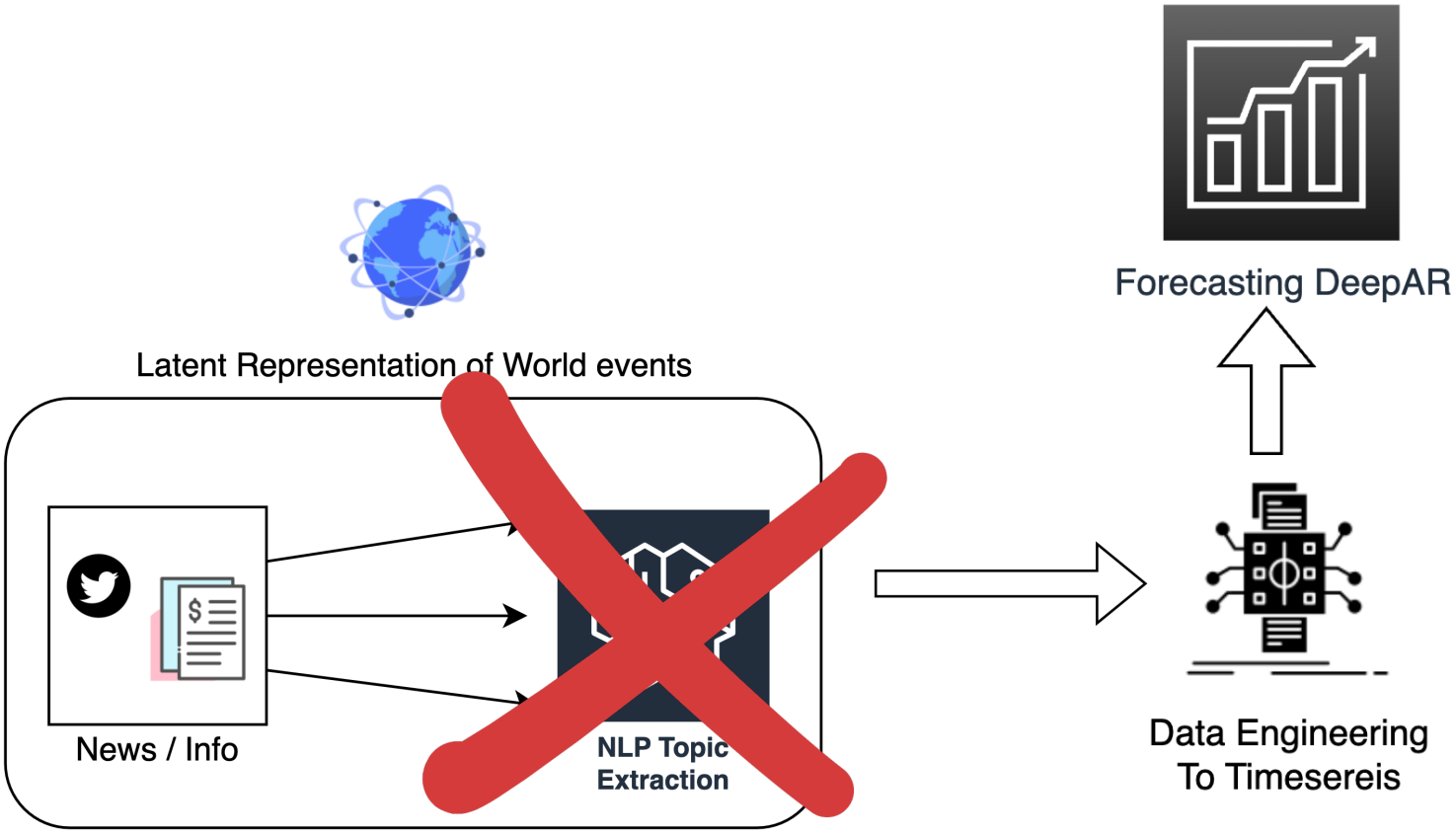
Modelling


Forecasting DeepAR
Latent Representation of World events
News / Info
NLP Topic
Extraction
Data Engineering
To Timesereis

Topic Modelling



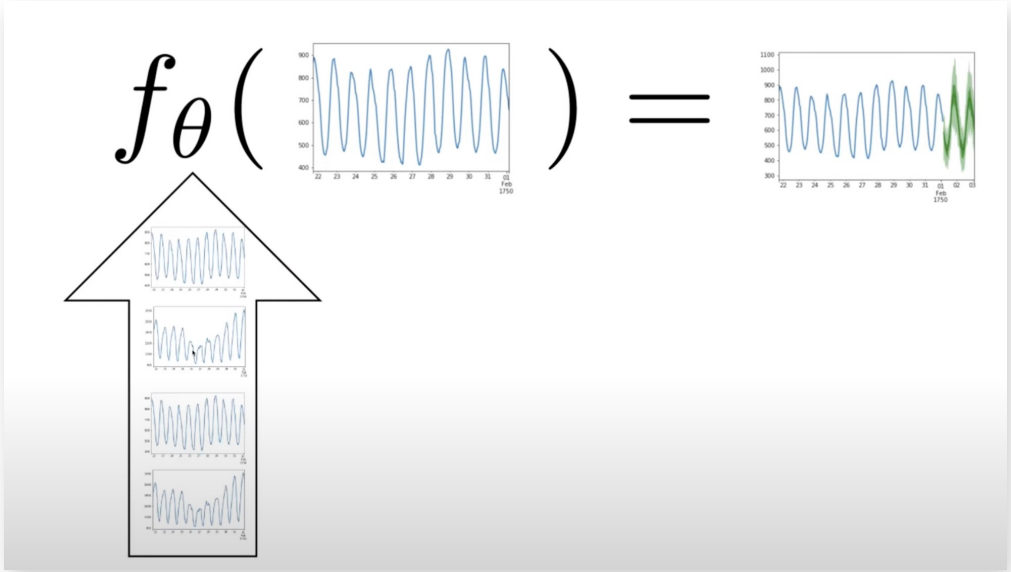
- Relevante Ereignisse aus News Artikeln extrahieren
- Quantitativer Statistischer Ansatz
  - Verwendung von Methoden wie LDA
  - Möglicherweise nicht akkurat für komplexe Artikel
  - geringer Rechenaufwand
- Qualitativer Deep Learning Ansatz
  - dynamische Themenerkennung
  - hoher Rechenaufwand


Forecasting DeepAR
Latent Representation of World events
News / Info
NLP Topic
Extraction
Data Engineering
To Timesereis


Latent Representation of World events
GDELT
Forecasting DeepAR
Data Engineering
To Timesereis

Forecasting

- Interpretation als global univariates Problem

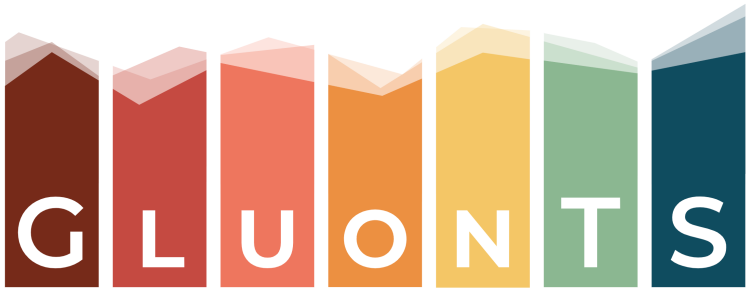


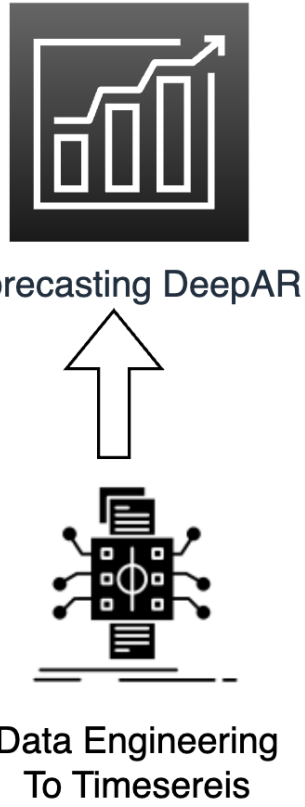
Forecasting DeepAR

Data Engineering
To Timesereis

Forecasting

- Interpretation als global univariates Problem
- Fokus auf komplexe Modelle aufgrund hoher Komplexität
  - MQ-CNN (CNN encoder, MLP decoder) - Wen et al. 2017
  - DeepAR (RNN) - Salinas et al. 2020
  - DeepState (RNN, state-space model) - Rangapuram et al. 2018

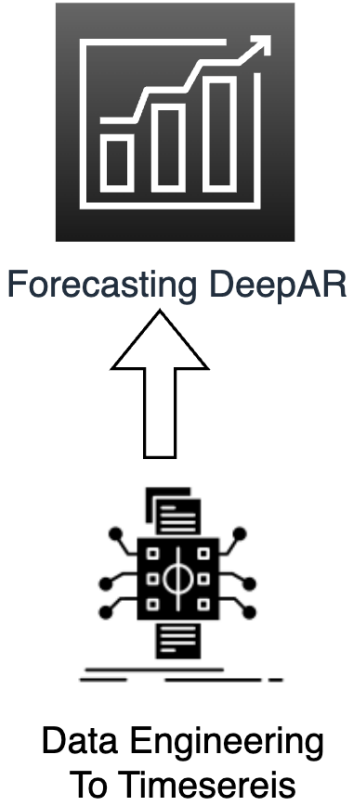




Forecasting

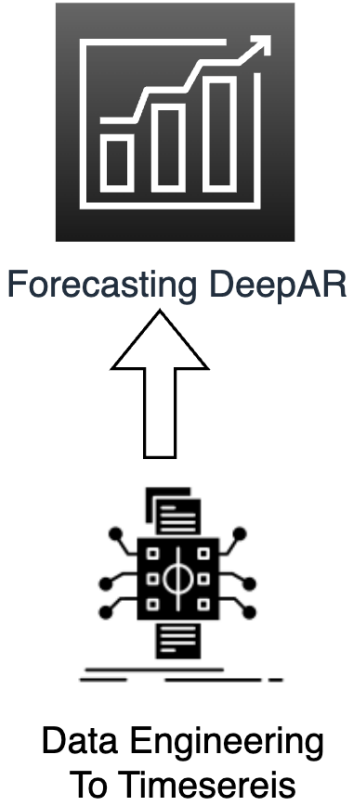
- Interpretation als global univariates Problem
- Fokus auf komplexe Modelle aufgrund hoher Komplexität
  - MQ-CNN (CNN encoder, MLP decoder) - Wen et al. 2017
  - DeepAR (RNN) - Salinas et al. 2020
  - DeepState (RNN, state-space model) - Rangapuram et al. 2018
- Hoher Rechenaufwand → Skalierbares Training in der Cloud

aws



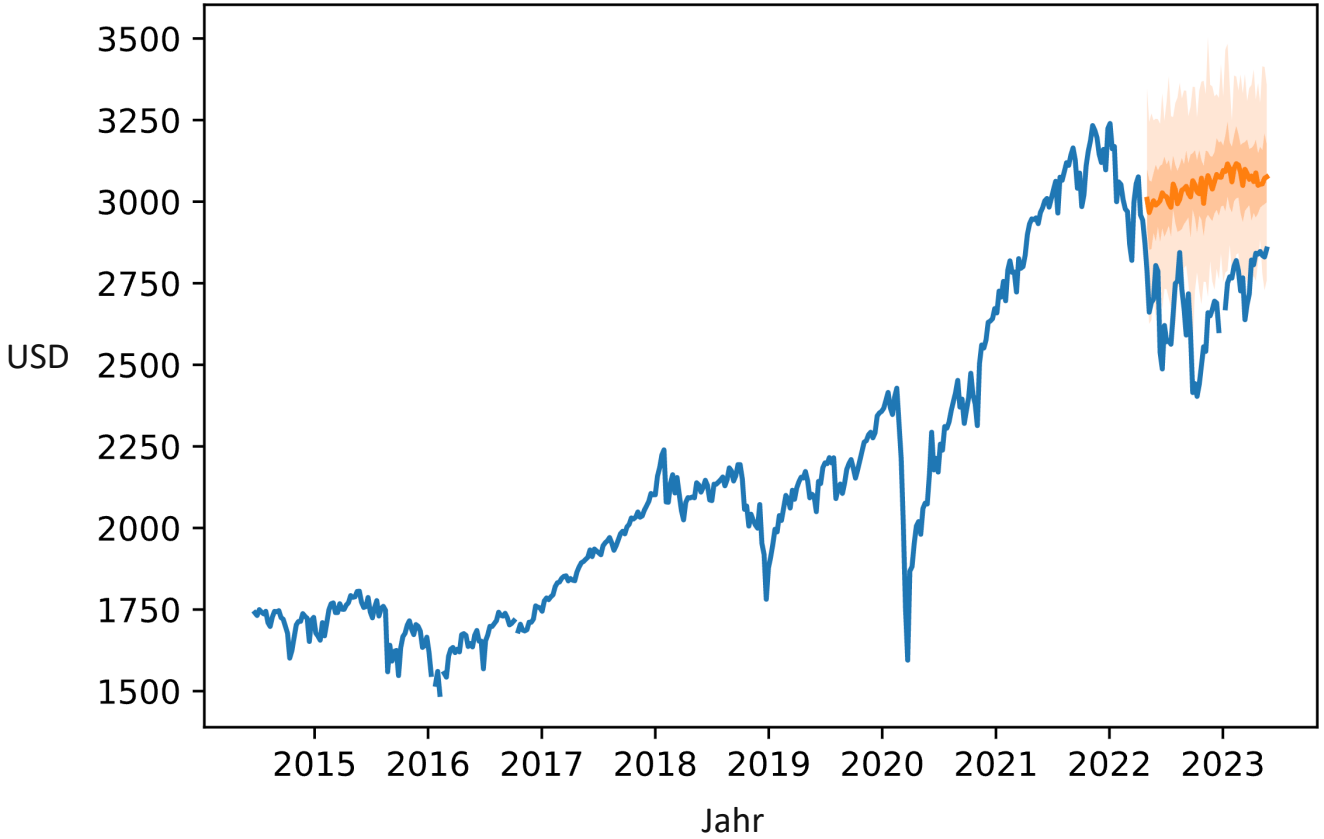
Forecasting

- Interpretation als global univariates Problem
- Fokus auf komplexe Modelle aufgrund hoher Komplexität
  - MQ-CNN (CNN encoder, MLP decoder) - Wen et al. 2017
  - DeepAR (RNN) - Salinas et al. 2020
  - DeepState (RNN, state-space model) - Rangapuram et al. 2018
- Hoher Rechenaufwand → Skalierbares Training in der Cloud
- Relationale Datenbank zur Dokumentation von Experimenten und deren Resultaten



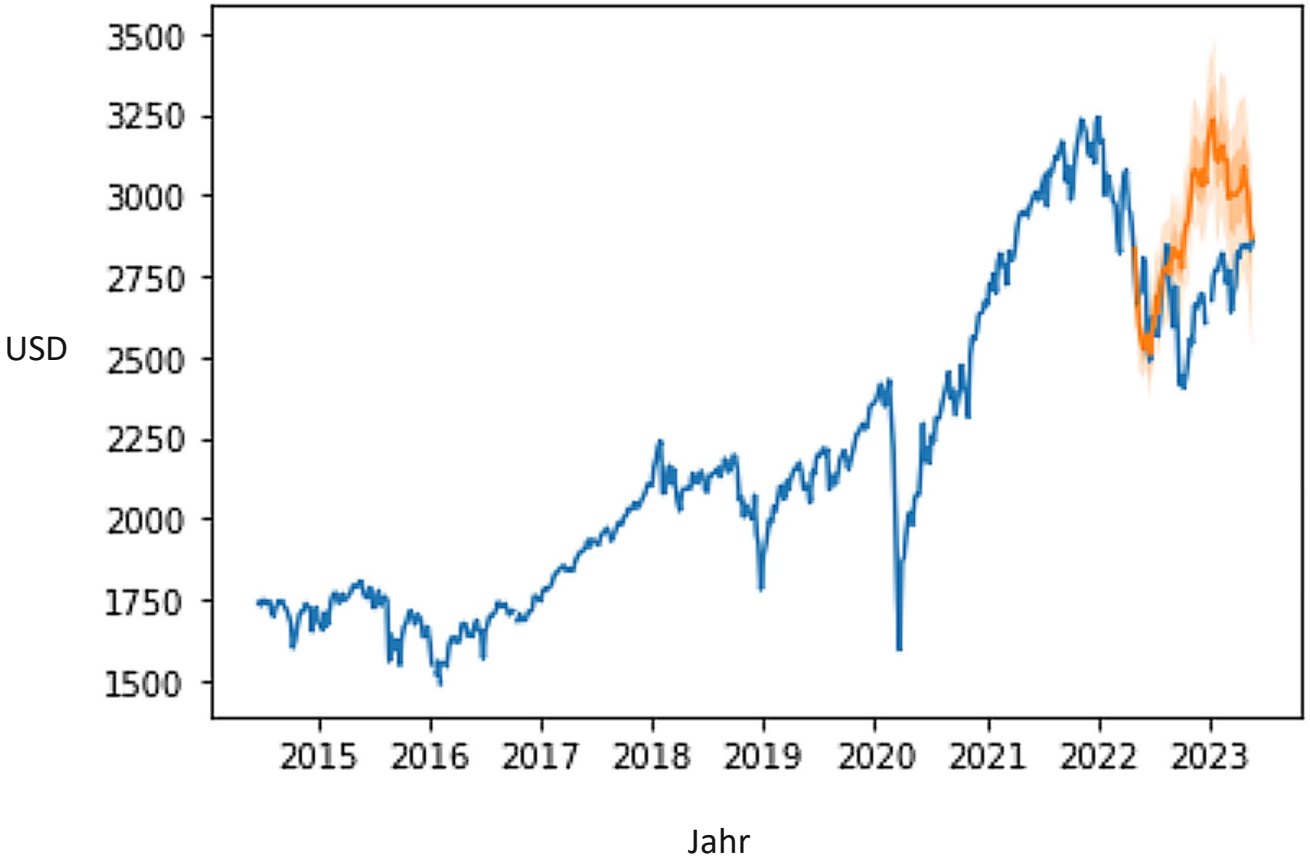
DeepAR

avg_WQL: 0.194



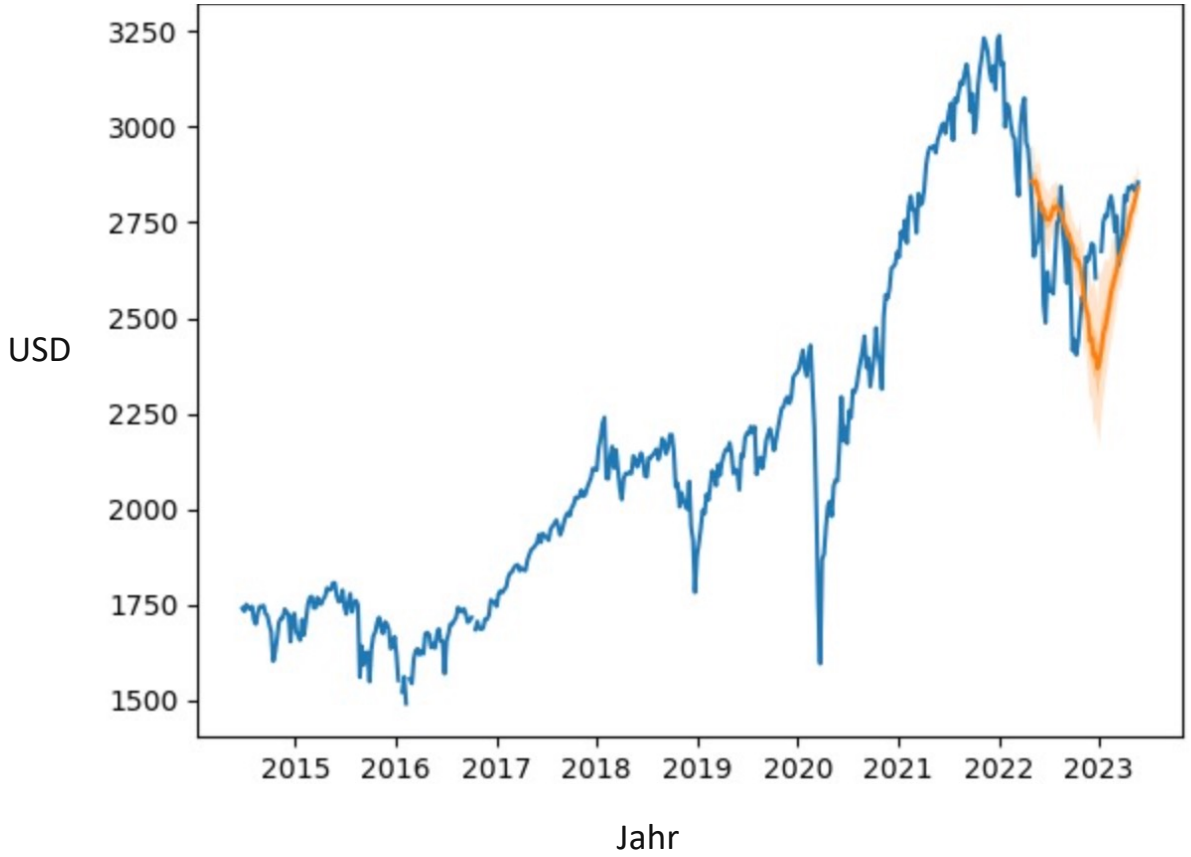
DeepState

avg_WQL: 0.082



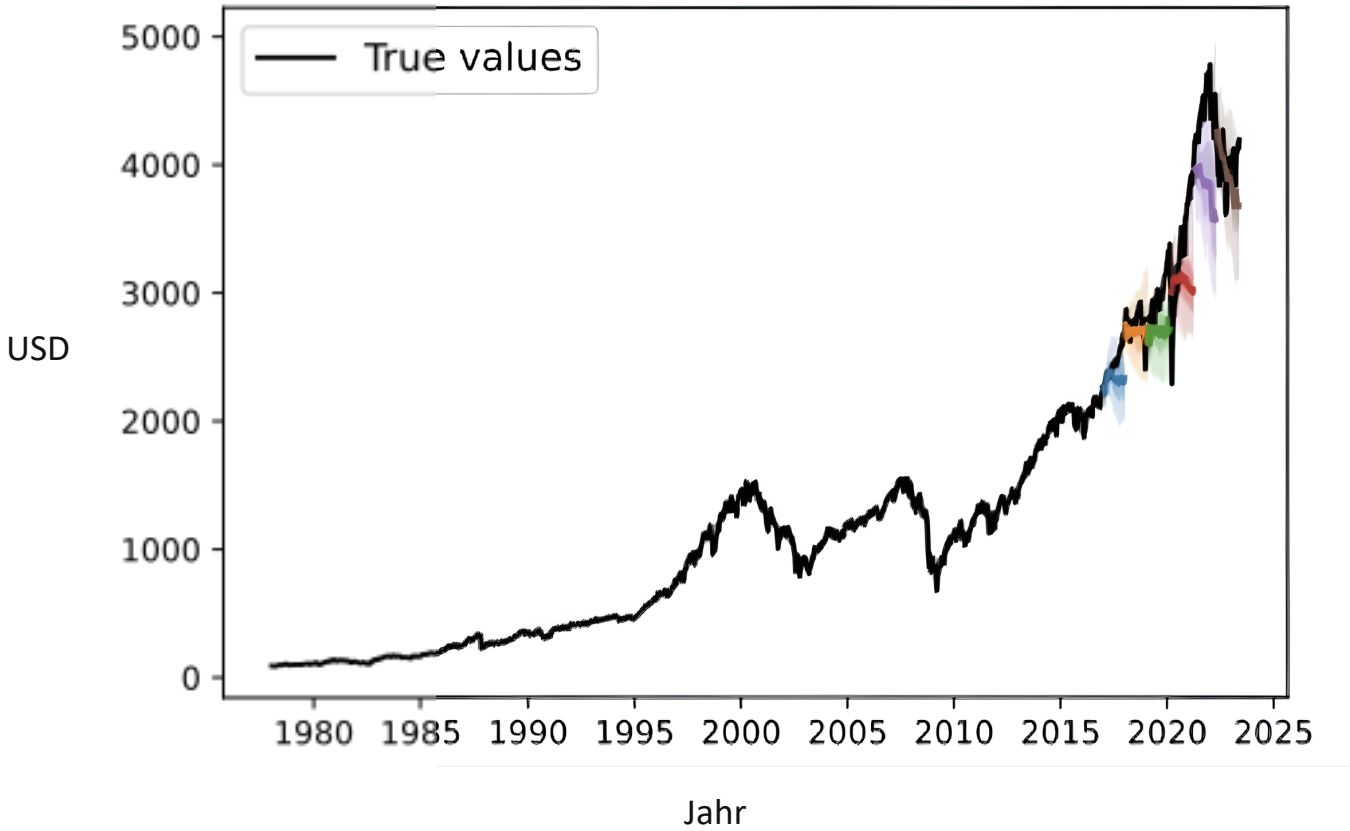
DeepAR

avg_WQL: 0.028



Rolling Backtest

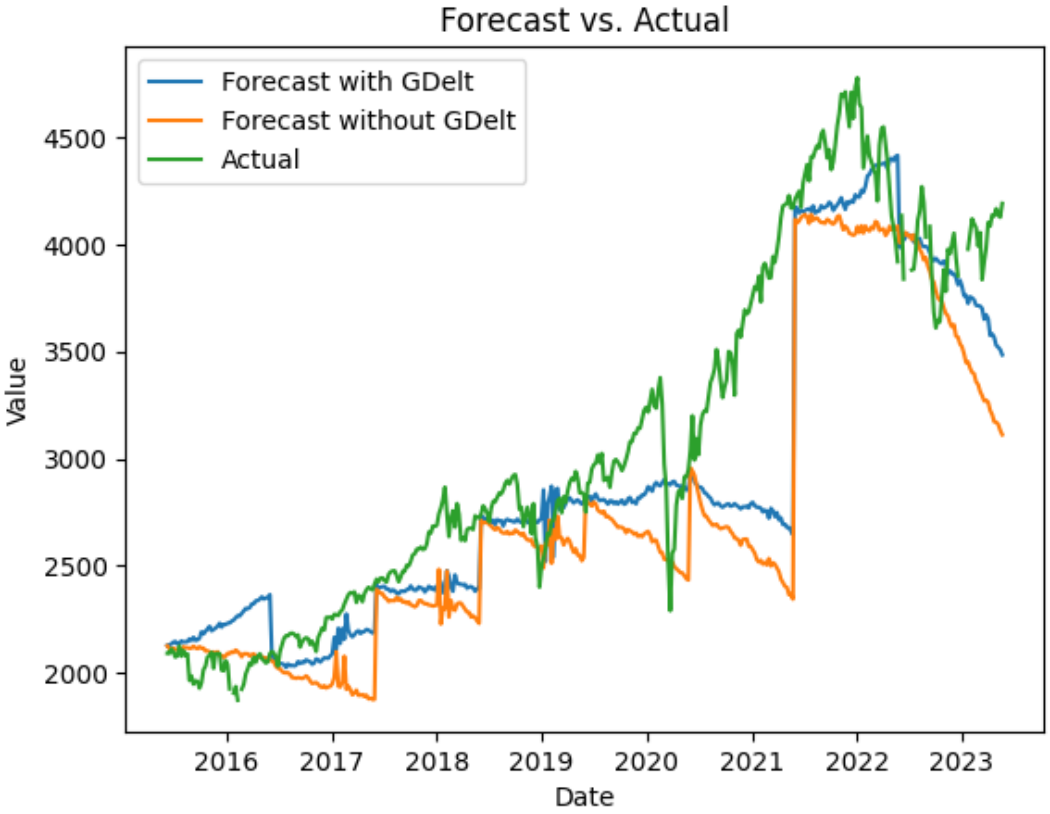
avg_WQL: 0.037



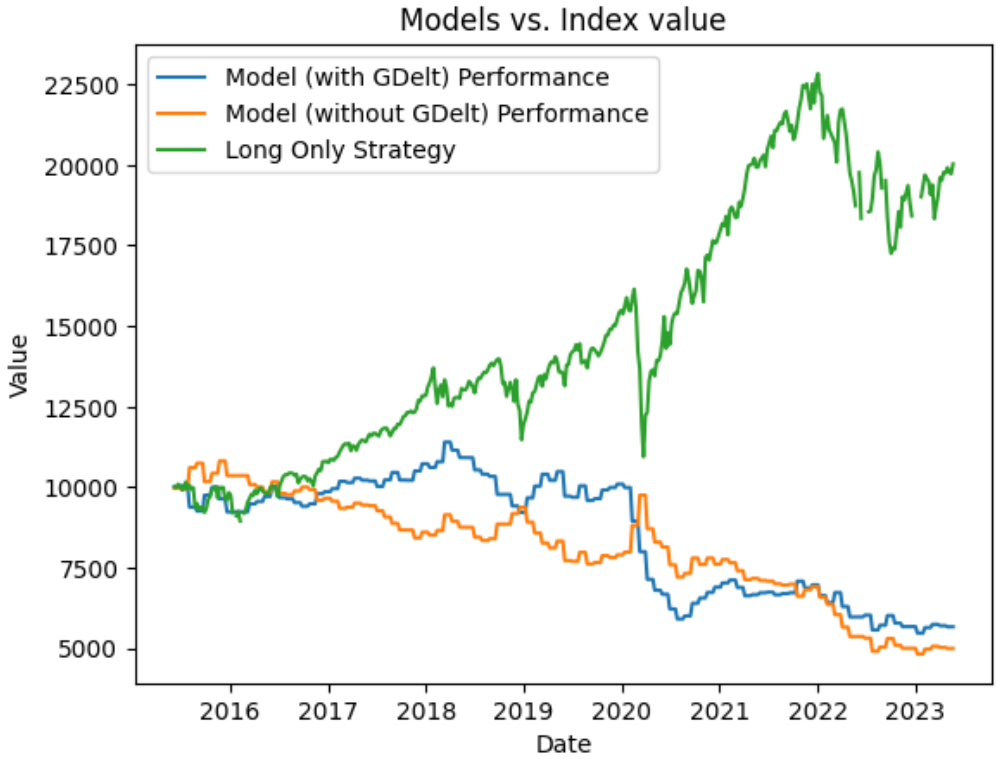
05

Ergebnis

Model Vorhersage mit/ohne GDELT

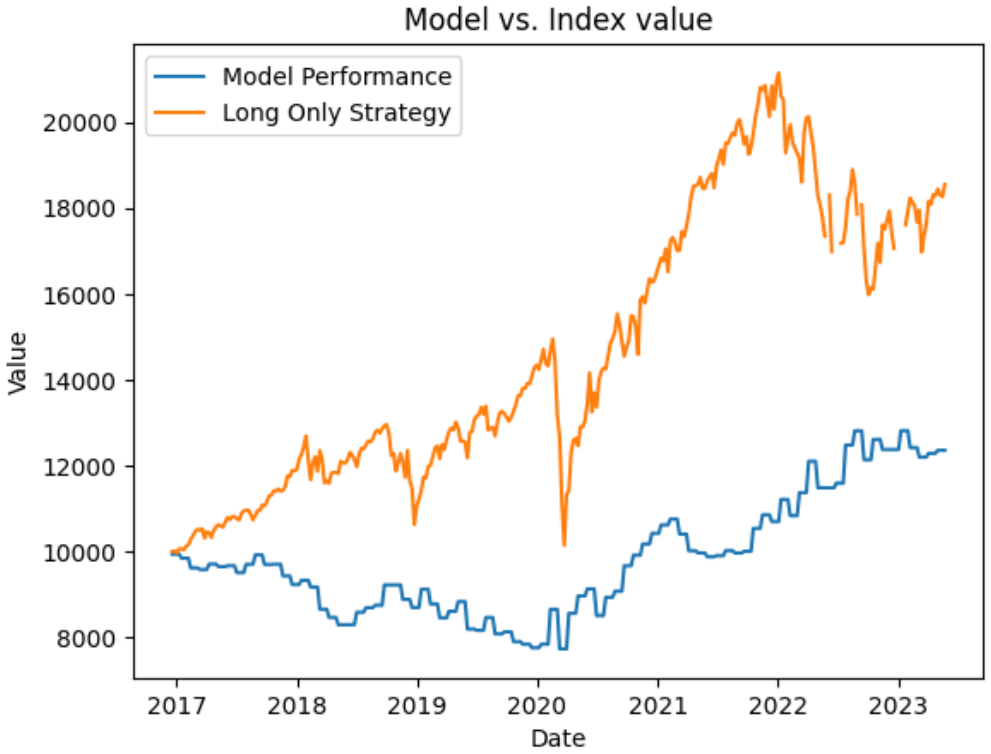


Model Performance mit/ohne GDELT



Model Performance

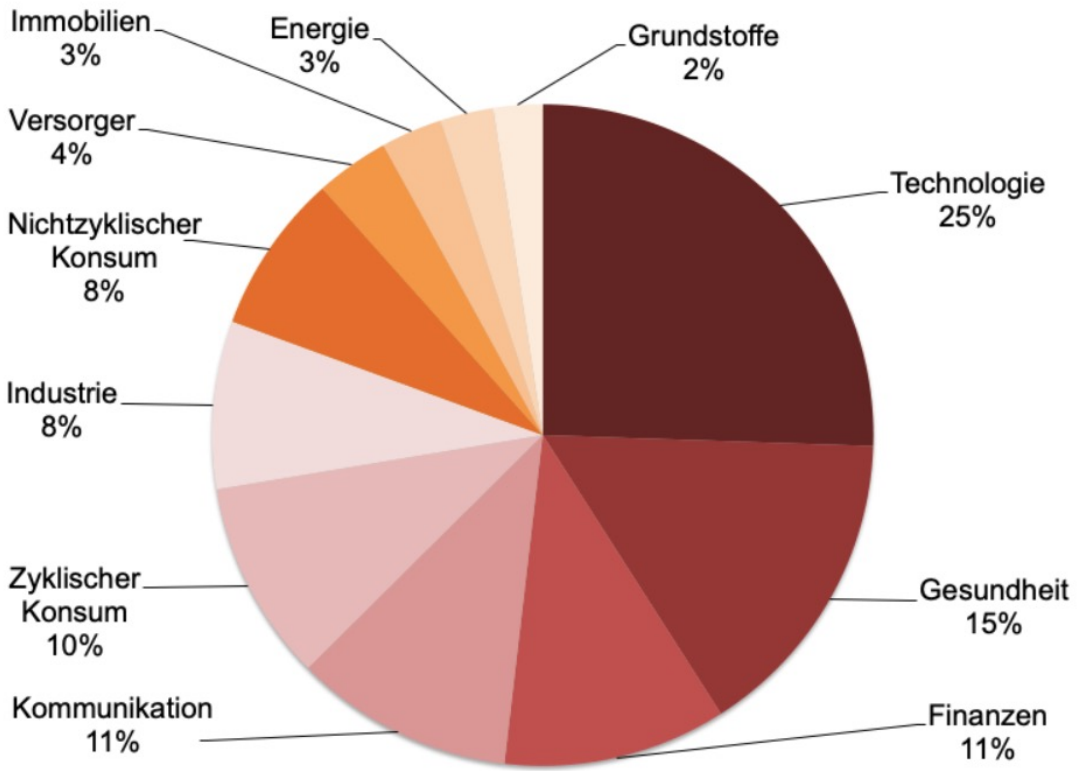
10.000€



Long only: 18.443€

Model: 12.335€

Aufbau S&P 500



Lessons learned

- **Daten visualisieren**
- **Investition in Datengrundlage**
- **Kleinschrittig Arbeiten**
- **Bessere Dokumentation**

Noch Fragen?

DeepState
ist ne
coole Sache

Damit
ist ein ML
Algorithmus
gemeint