

Zwei-Stufen-XGBoost - Experiment-Report

Experiment-ID: v1_h4_thr0p5pct_strict

Dieses Dokument fasst die wichtigsten Parameter, Datenquellen und Metriken eines Zwei-Stufen-XGBoost-Experiments zusammen.

Stufe 1 (Signal): neutral vs. Bewegung ('move'). Stufe 2 (Richtung): down vs. up – nur an Bewegungstagen.

Label-Parameter:

- horizon_days: 4
- up_threshold: 0.005
- down_threshold: -0.005
- strict_monotonic: True

Datensatz & Splits:

- dataset_path: /Users/jeremynathan/Documents/GitHub/hs2025_ml_project/hs2025_ml_project/data/processed/datasets/eurusd_news_training.csv
- test_start: 2025-01-01
- train_frac_within_pretest: 0.8

Features (FEATURE_COLS): vollständige Liste auf der Feature-Seite weiter unten.

Legende & Begriffe

Zielvariablen:

- label: 3-Klassen-Ziel auf Basis des 4-Tage-Lookaheads (neutral / up / down).
- signal: 0 = neutral, 1 = Bewegung (up oder down).
- direction: 0 = down, 1 = up; nur definiert, wenn signal == 1.

Wichtige Metriken:

- precision: Anteil der vorhergesagten positiven Fälle, die wirklich positiv sind.
- recall: Anteil der tatsächlichen positiven Fälle, die erkannt wurden.
- f1: harmonischer Mittelwert aus precision und recall (Balance beider Größen).
- support: Anzahl der Beobachtungen in der jeweiligen Klasse.

Feature-Abkürzungen (Auswahl):

- article_count: Anzahl News-Artikel pro Tag.
- avg_polarity / avg_neg / avg_neu / avg_pos: durchschnittliche Sentiment-Werte.
- pos_share / neg_share: Anteil positiver bzw. negativer Sentiment-Komponente.
- intraday_range_pct: $(\text{High} - \text{Low}) / \text{Close}$ – relative Tages-Spanne (Volatilität).
- upper_shadow / lower_shadow: obere/untere Dochte der Kerzen (High/Low vs. Körper).
- month / quarter: Kalendermonat und Quartal.

Modell-Parameter (XGBoost)

Signal-Modell (Stufe 1):

- objective: binary:logistic
- max_depth: 3
- learning_rate: 0.05
- n_estimators: None
- subsample: 0.9
- colsample_bytree: 0.9
- scale_pos_weight: 7.689655172413793

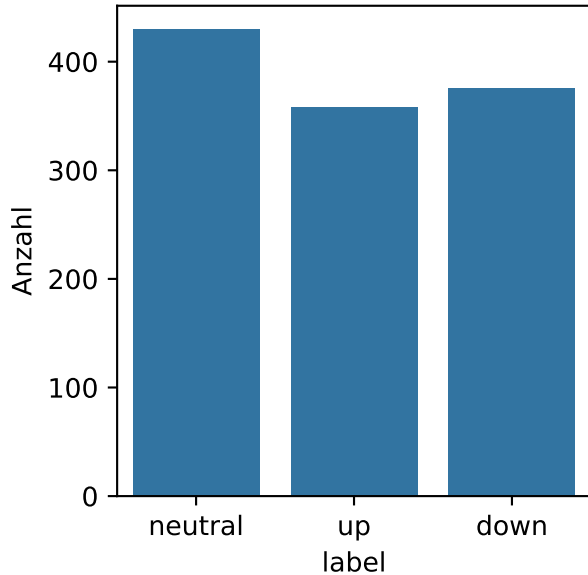
Richtungs-Modell (Stufe 2):

- objective: binary:logistic
- max_depth: 3
- learning_rate: 0.05
- n_estimators: None
- subsample: 0.9
- colsample_bytree: 0.9
- scale_pos_weight: 1.0

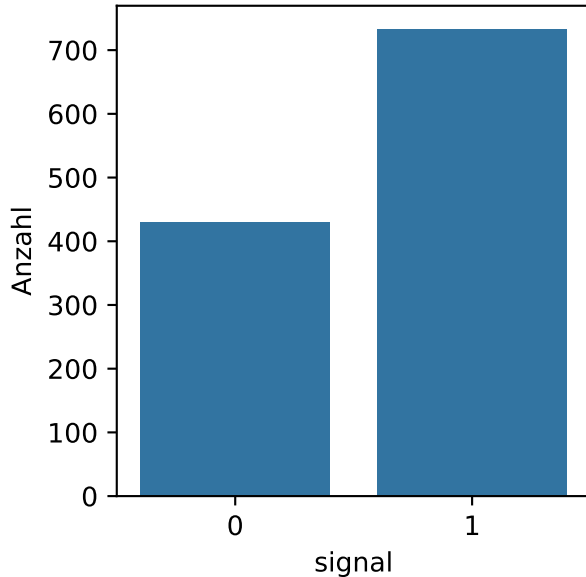
Verwendete Features (FEATURE_COLS)

#	feature_name
0	article_count
1	avg_polarity
2	avg_neg
3	avg_neu
4	avg_pos
5	pos_share
6	neg_share
7	intraday_range_pct
8	upper_shadow
9	lower_shadow
10	month
11	quarter

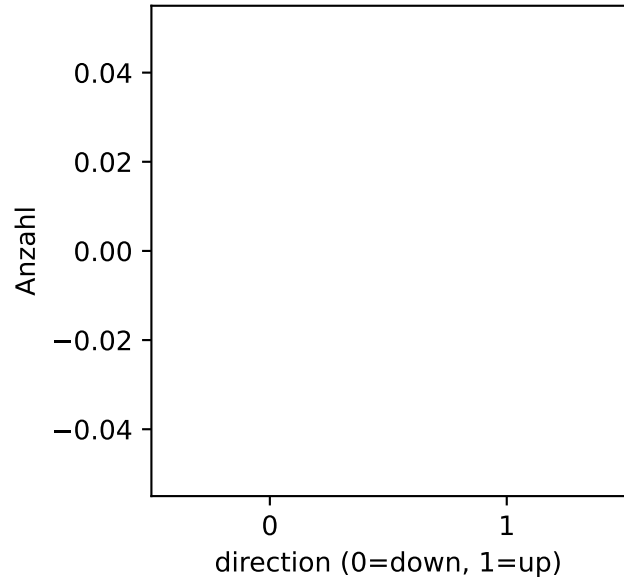
Label-Verteilung (neutral / up / down)



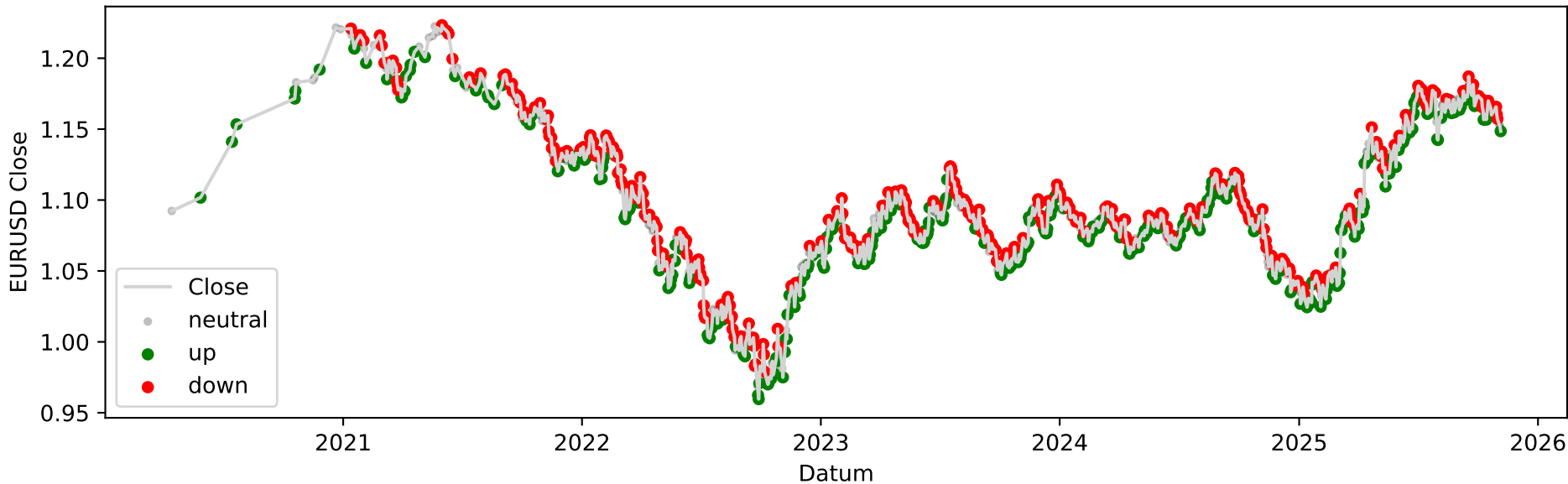
Signal-Verteilung (0=neutral, 1=move)



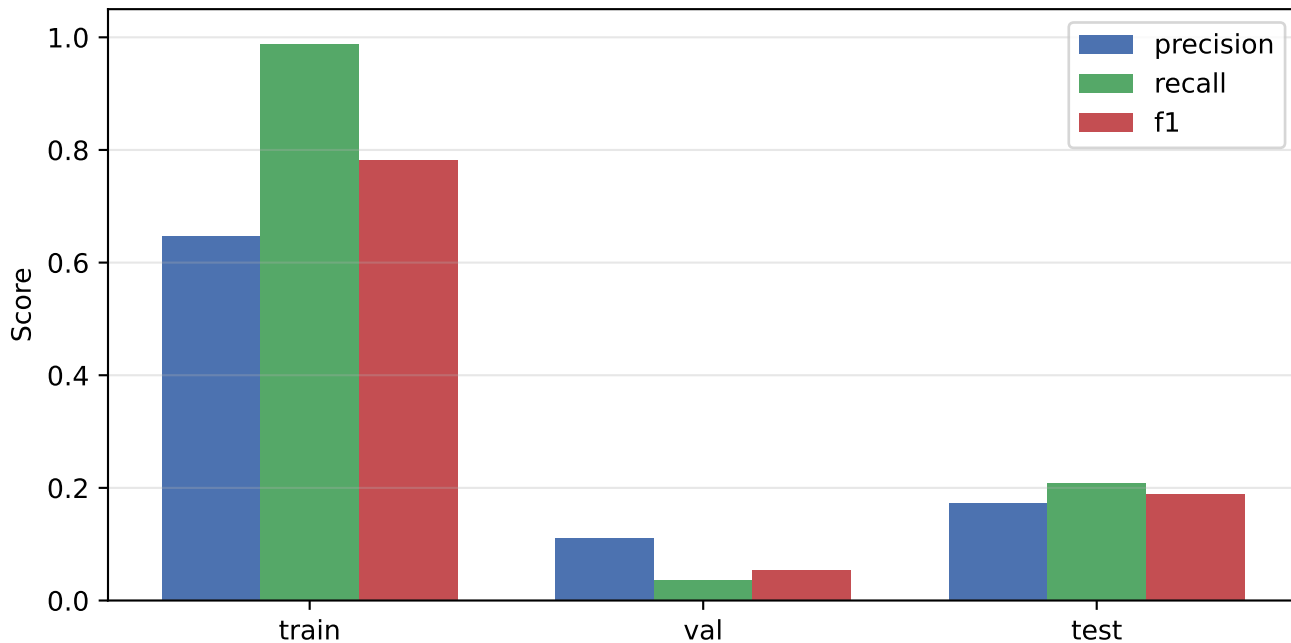
Richtung-Verteilung (nur signal==1)



EURUSD-Zeitreihe mit hervorgehobenen up/down-Tagen (ab 2020)



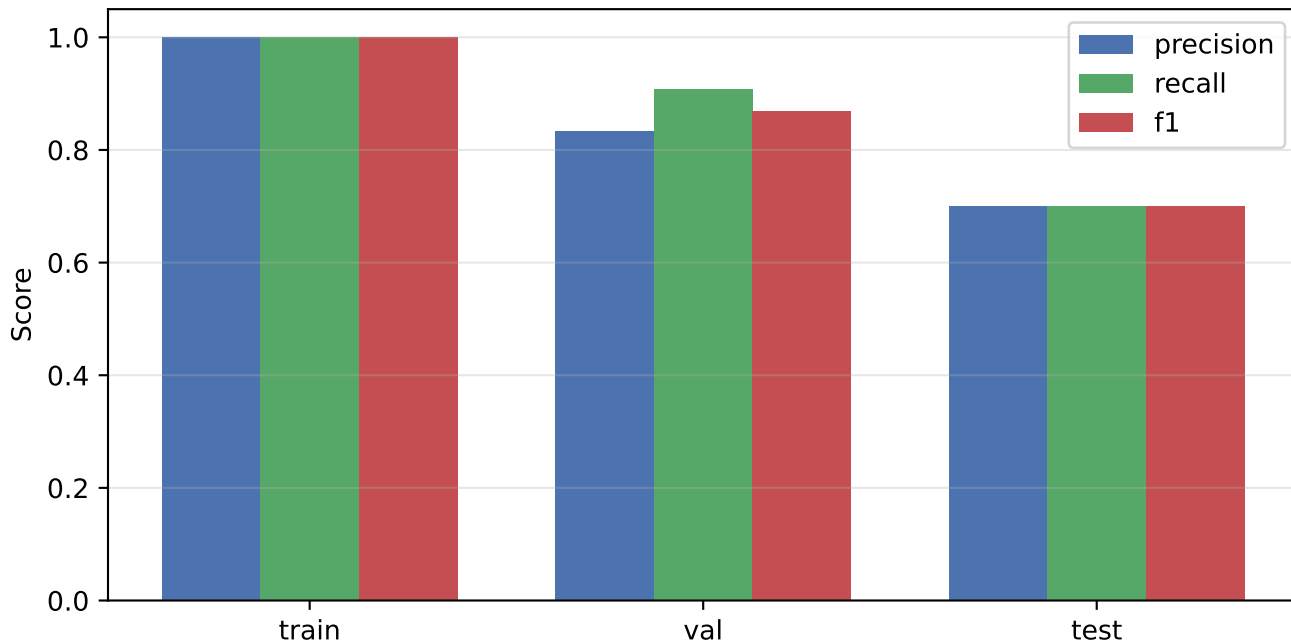
Signal-Modell - Kennzahlen für Klasse 'move' (train/val/test)



Signal-Modell - Tabelle (Klasse 'move')

split	precision	recall	f1	support
train	0.647	0.989	0.782	87.000
val	0.111	0.036	0.054	28.000
test	0.172	0.208	0.189	24.000

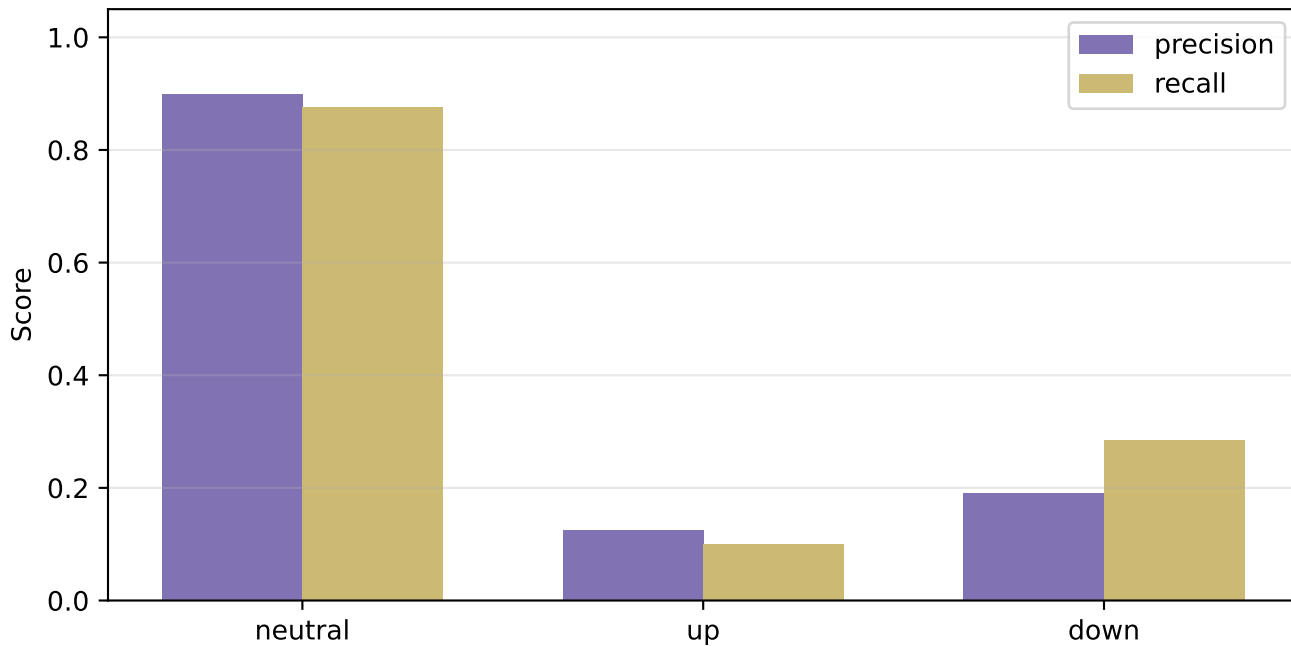
Richtungs-Modell - Kennzahlen für Klasse 'up' (train/val/test)



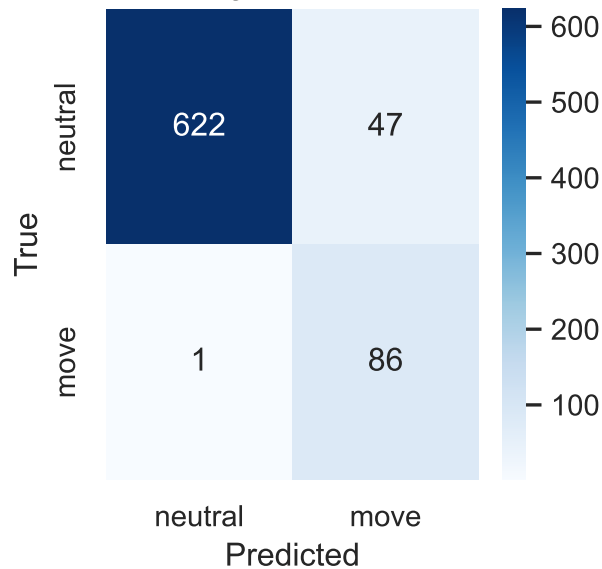
Richtungs-Modell - Tabelle (Klasse 'up')

split	precision	recall	f1	support
train	1.000	1.000	1.000	38.000
val	0.833	0.909	0.870	11.000
test	0.700	0.700	0.700	10.000

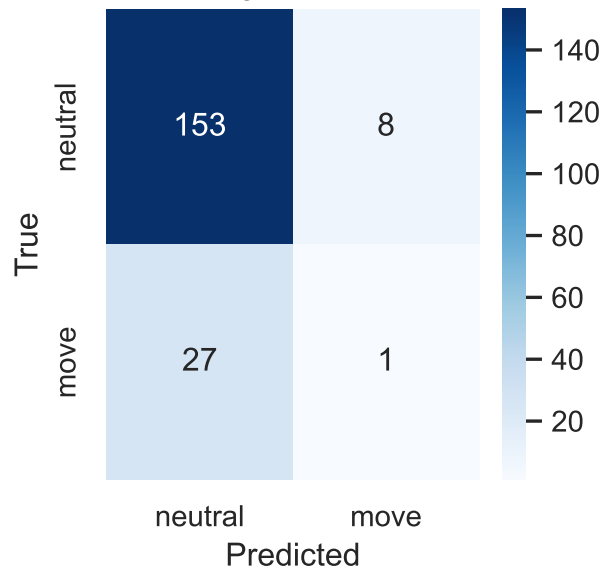
Kombinierte Test-Auswertung - neutral / up / down



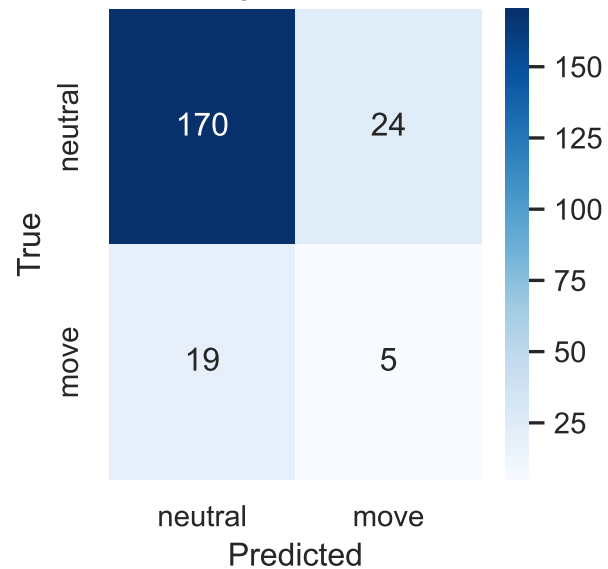
Signal – train



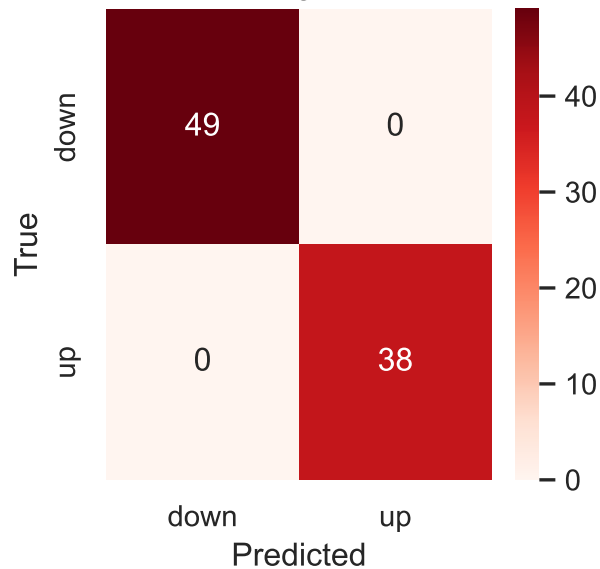
Signal – val



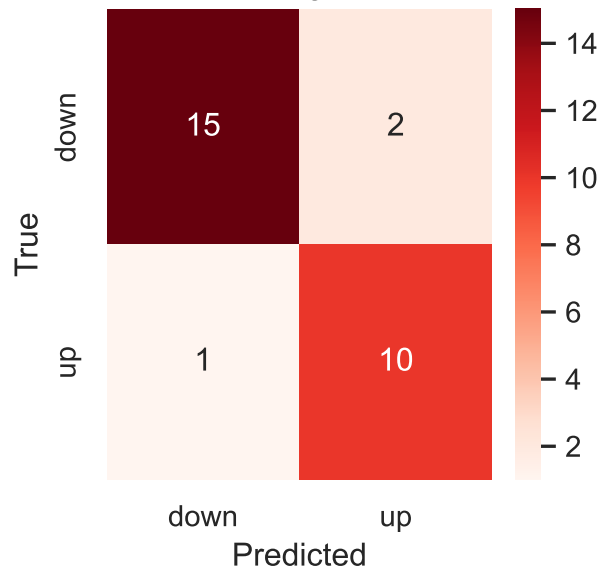
Signal – test



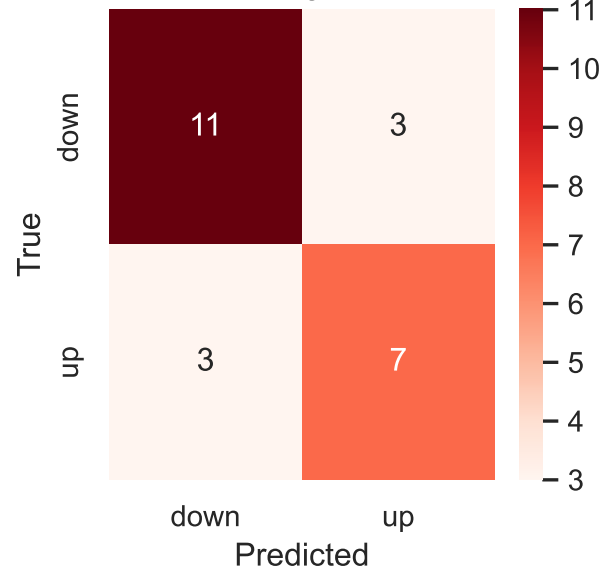
Richtung – train



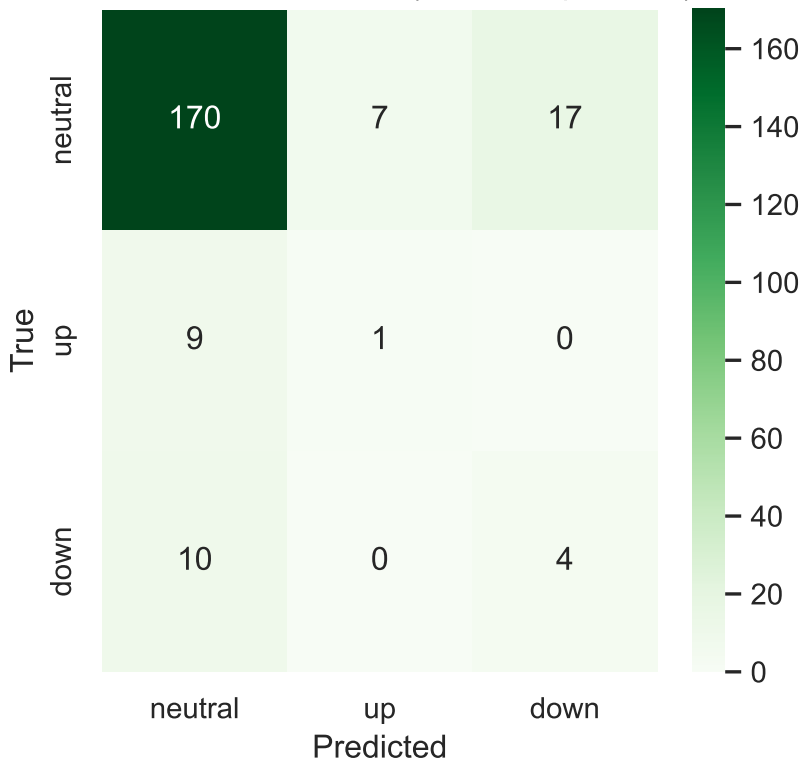
Richtung – val



Richtung – test



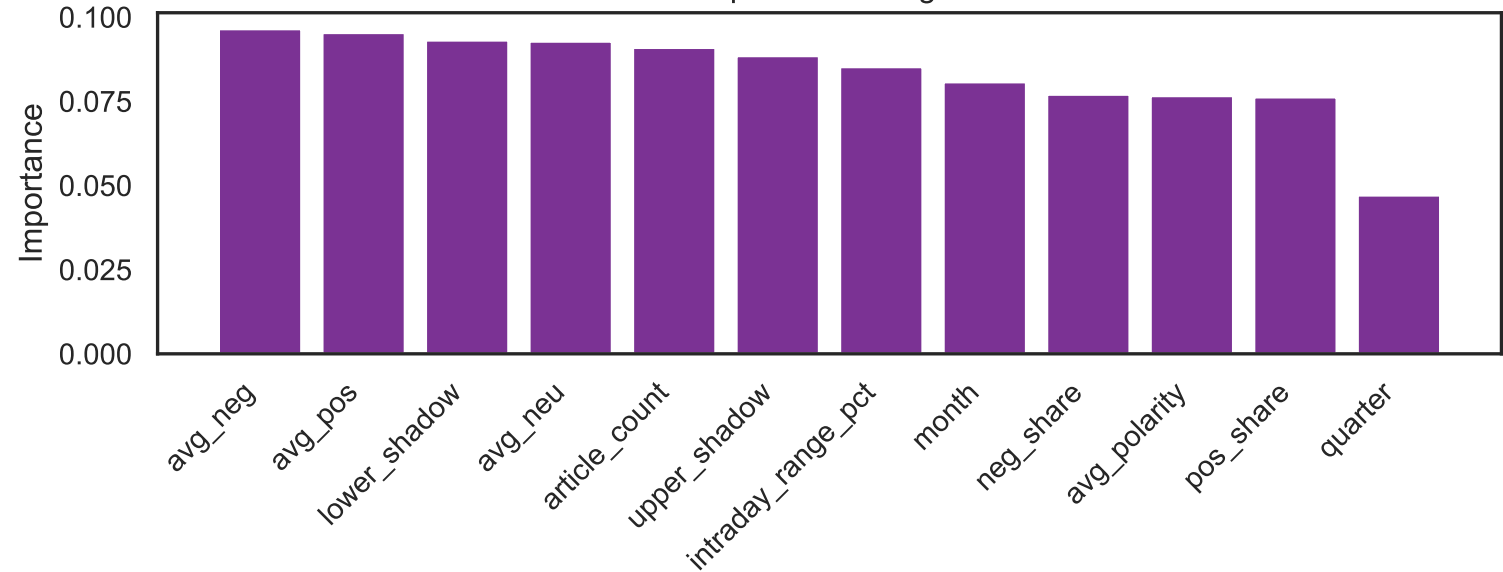
Confusion Matrix – Test (neutral / up / down)



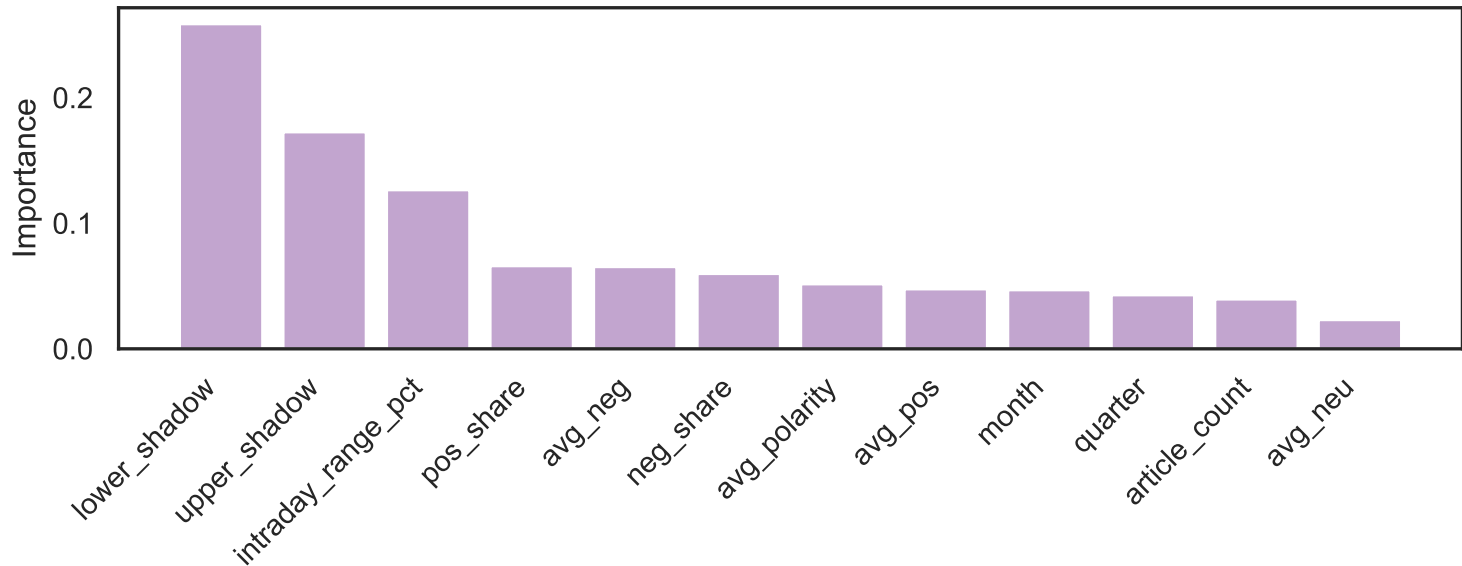
Konfusionsmatrizen – Zählwerte (TN/FP/FN/TP)

modell	split	TN	FP	FN	TP
signal (neutral vs. mov	train	622	47	1	86
signal (neutral vs. mov	val	153	8	27	1
signal (neutral vs. mov	test	170	24	19	5
direction (down vs. up	train	49	0	0	38
direction (down vs. up	val	15	2	1	10
direction (down vs. up	test	11	3	3	7

Feature Importance – Signal-Modell



Feature Importance – Richtungs-Modell



Confusion Matrix – Test (Richtung: down vs. up)

