

Zwei-Stufen-XGBoost - Experiment-Report

Experiment-ID: v7_h4_thr0p5pct_tolerant0p3pct_sigdepth4_n600

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: False

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: 4
- learning_rate: 0.05
- n_estimators: None
- subsample: 0.9
- colsample_bytree: 0.9
- scale_pos_weight: 1.2366863905325445

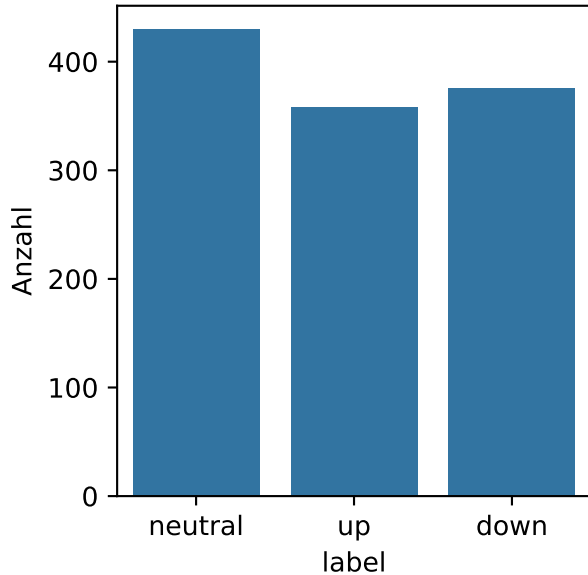
Richtungs-Modell (Stufe 2):

- 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

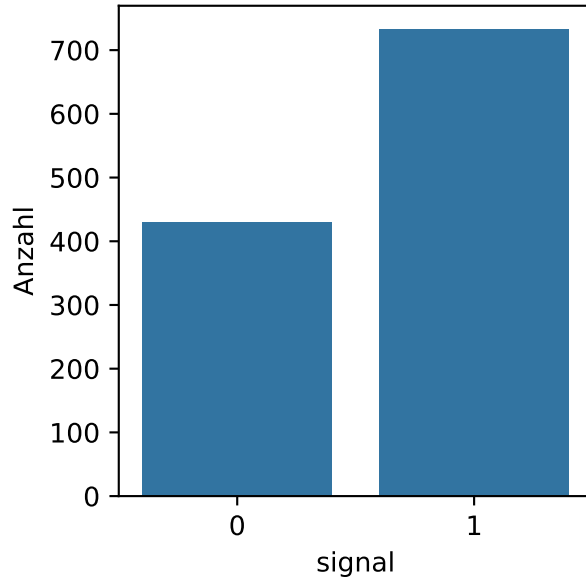
Verwendete Features (FEATURE_COLS)

#	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	price_close_ret_1d
11	price_close_ret_5d
12	price_range_pct_5d_std
13	price_body_pct_5d_mean
14	news_article_count_3d_sum
15	news_article_count_7d_sum
16	news_pos_share_5d_mean
17	news_neg_share_5d_mean
18	news_article_count_lag1
19	news_pos_share_lag1
20	news_neg_share_lag1
21	month
22	quarter
23	cal_dow
24	cal_day_of_month
25	cal_is_monday
26	cal_is_friday
27	cal_is_month_start
28	cal_is_month_end
29	hol_is_us_federal_holiday
30	hol_is_day_before_us_federal_holiday
31	hol_is_day_after_us_federal_holiday

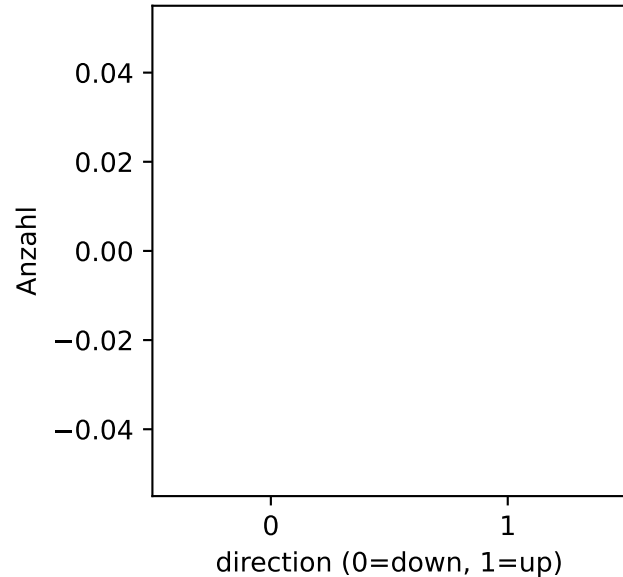
Label-Verteilung (neutral / up / down)



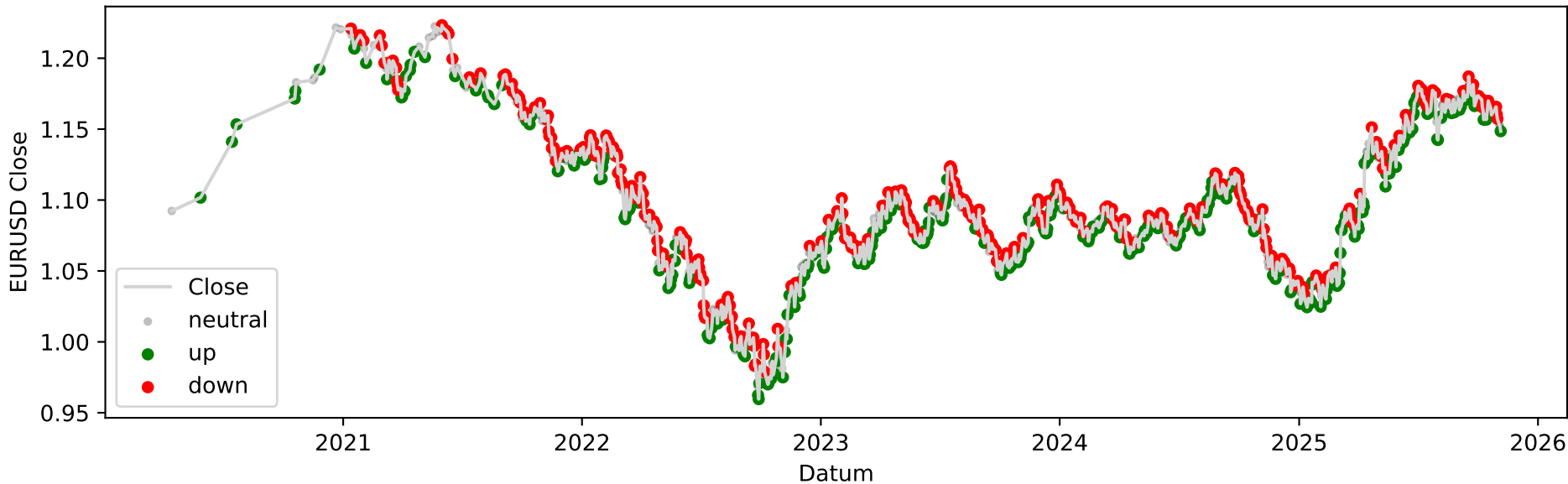
Signal-Verteilung (0=neutral, 1=move)



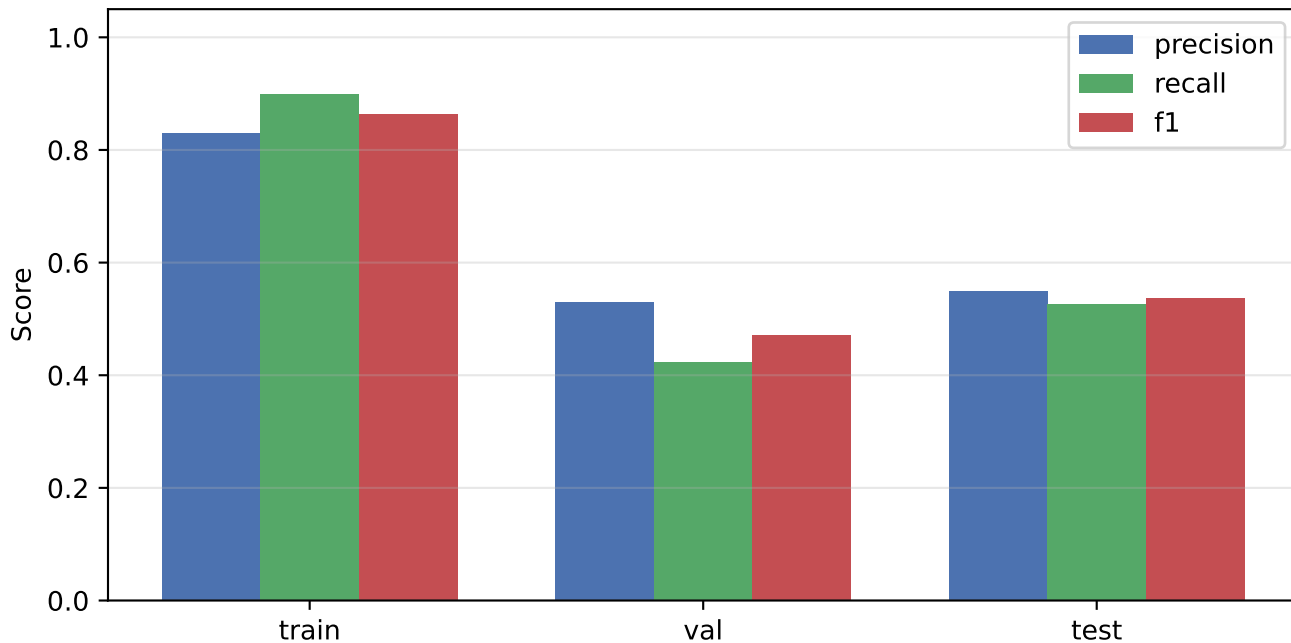
Richtung-Verteilung (nur signal==1)



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



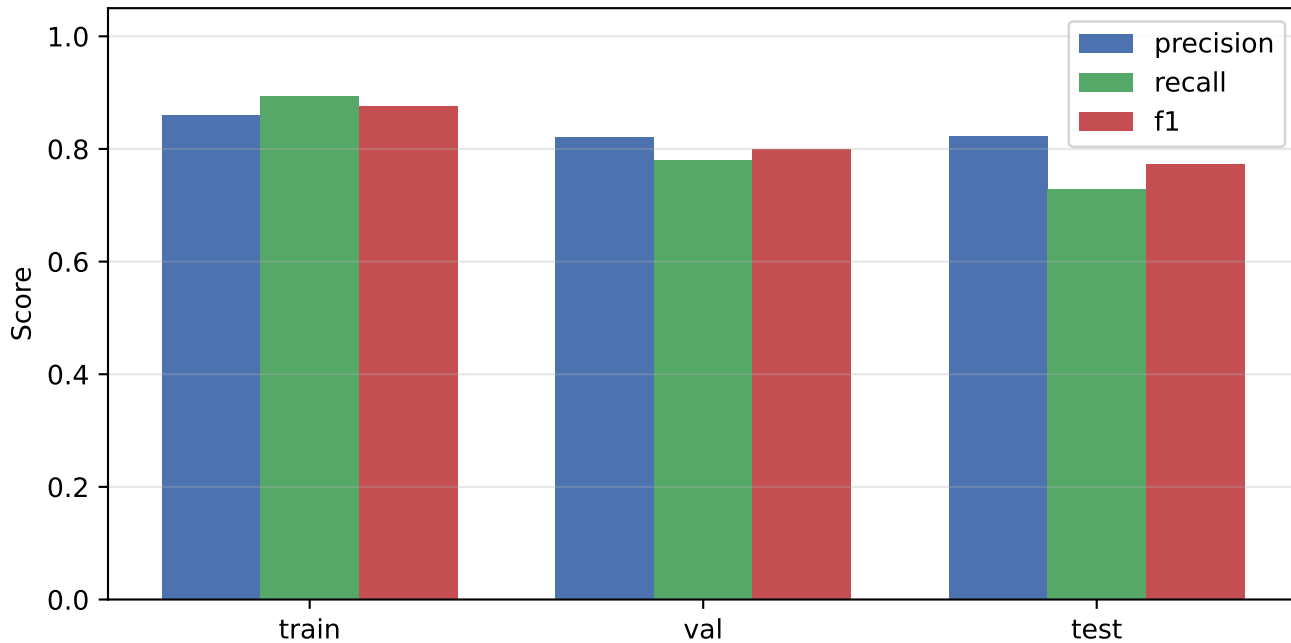
Signal-Modell - Kennzahlen für Klasse 'move' (train/val/test)



Signal-Modell - Tabelle (Klasse 'move')

split	precision	recall	f1	support
train	0.831	0.899	0.864	338.000
val	0.529	0.424	0.471	85.000
test	0.550	0.526	0.537	116.000

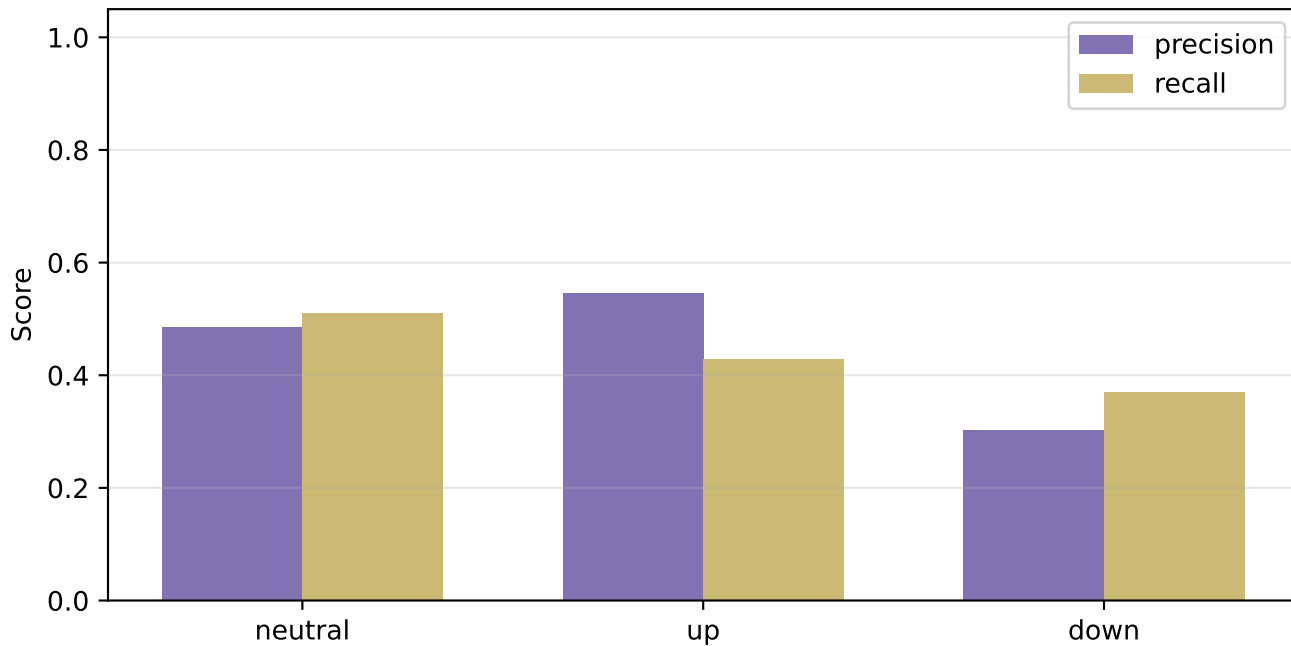
Richtungs-Modell - Kennzahlen für Klasse 'up' (train/val/test)

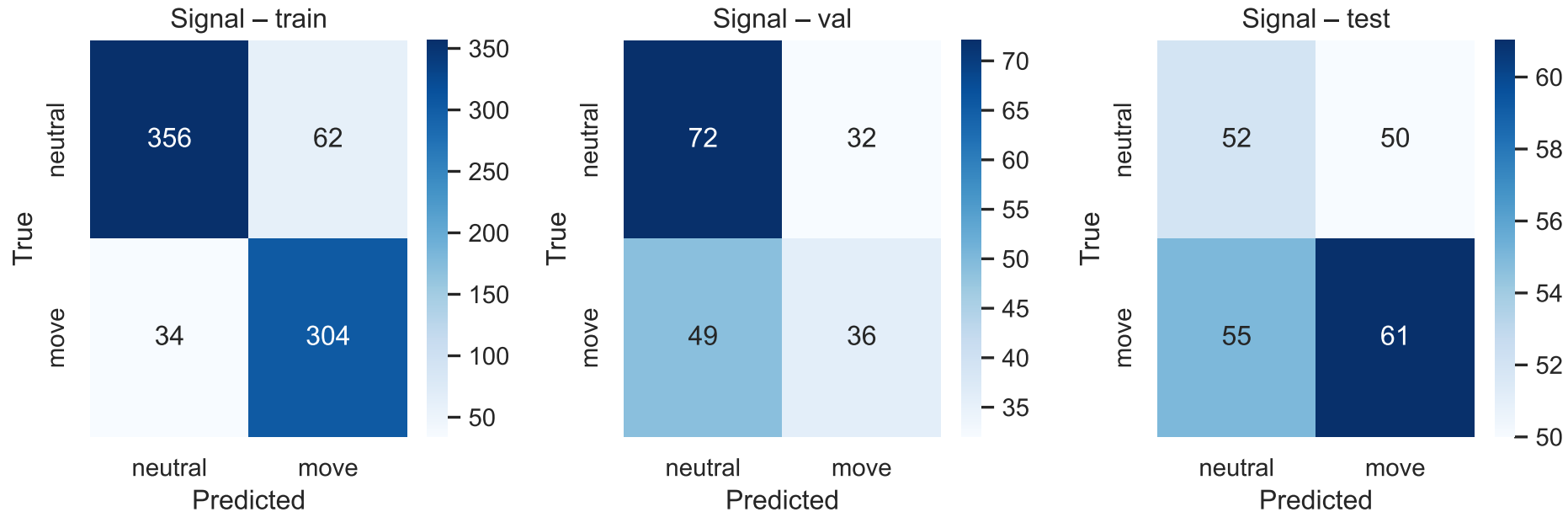


Richtungs-Modell - Tabelle (Klasse 'up')

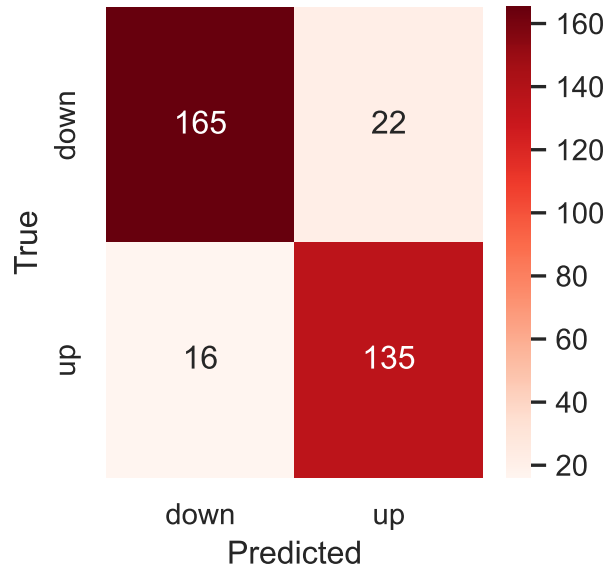
split	precision	recall	f1	support
train	0.860	0.894	0.877	151.000
val	0.821	0.780	0.800	41.000
test	0.823	0.729	0.773	70.000

Kombinierte Test-Auswertung - neutral / up / down

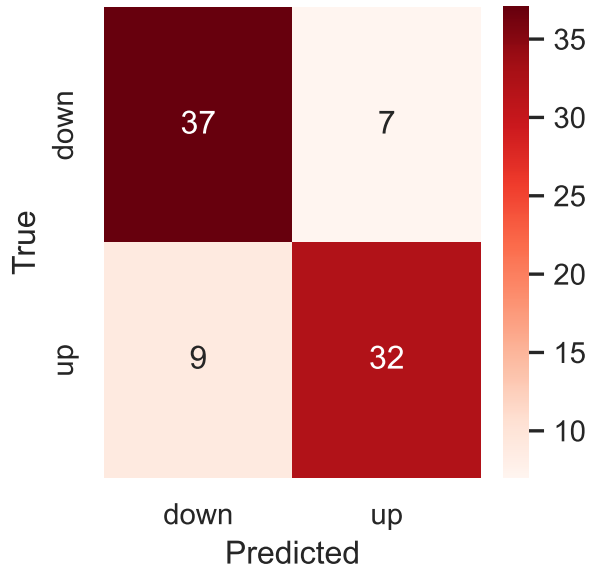




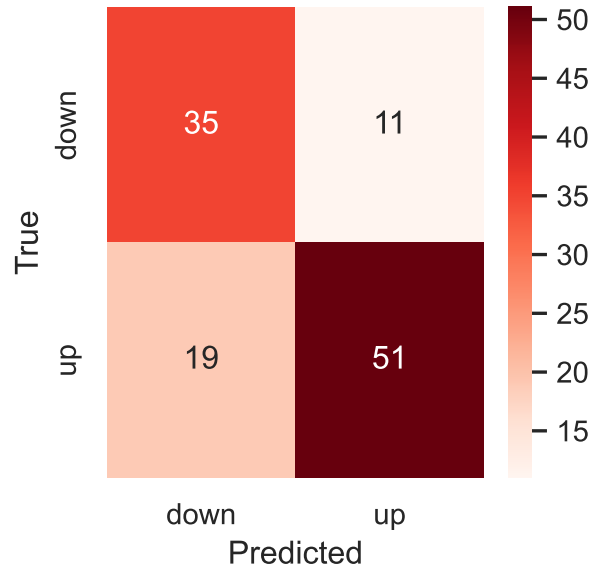
Richtung – train



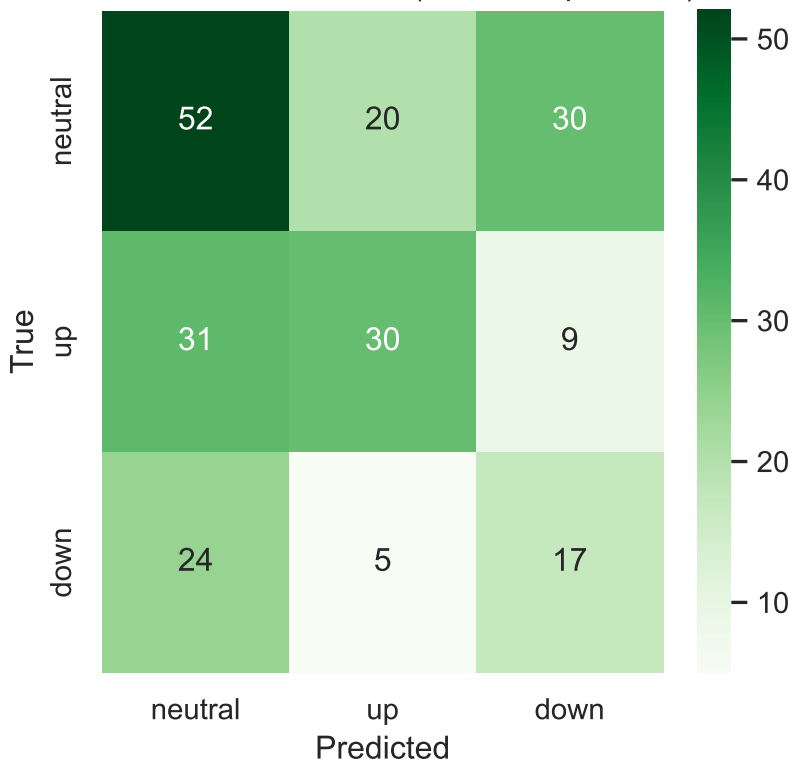
Richtung – val



Richtung – test



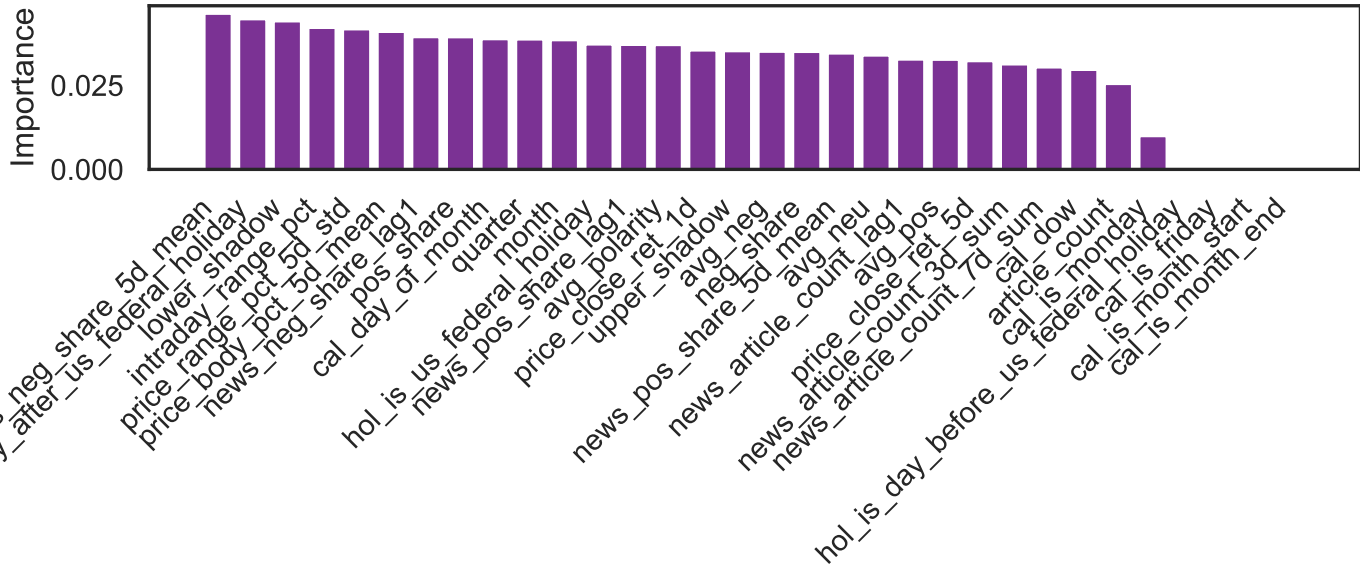
Confusion Matrix – Test (neutral / up / down)



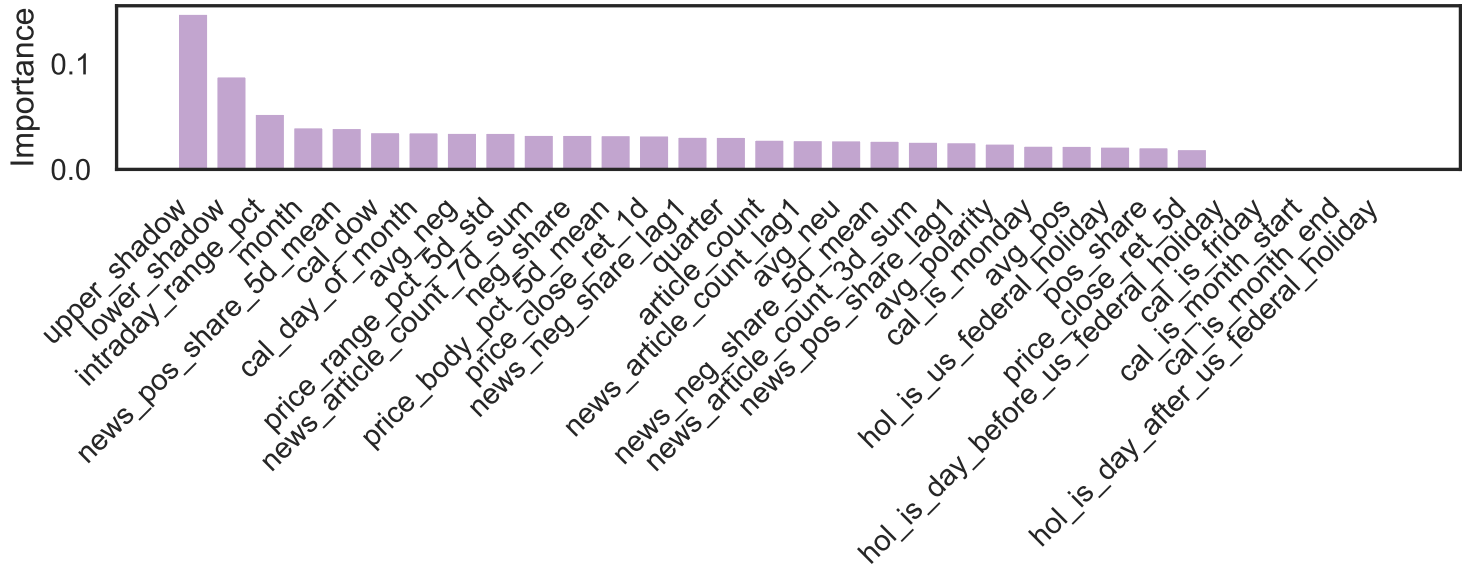
Konfusionsmatrizen – Zählwerte (TN/FP/FN/TP)

modell	split	TN	FP	FN	TP
signal (neutral vs. mov	train	356	62	34	304
signal (neutral vs. mov	val	72	32	49	36
signal (neutral vs. mov	test	52	50	55	61
direction (down vs. up	train	165	22	16	135
direction (down vs. up	val	37	7	9	32
direction (down vs. up	test	35	11	19	51

Feature Importance – Signal-Modell



Feature Importance – Richtungs-Modell



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

