MACHINE LEARNING FOR FOREX FORECASTING

Kevin Jeswani
WCD Bootcamp, 2022 (Part-Time)

Overview

- ML in quantitative finance
- Data gathering, cleaning
- EDA
- Classical forecasting
- Boosted Trees
- RNN - LSTM

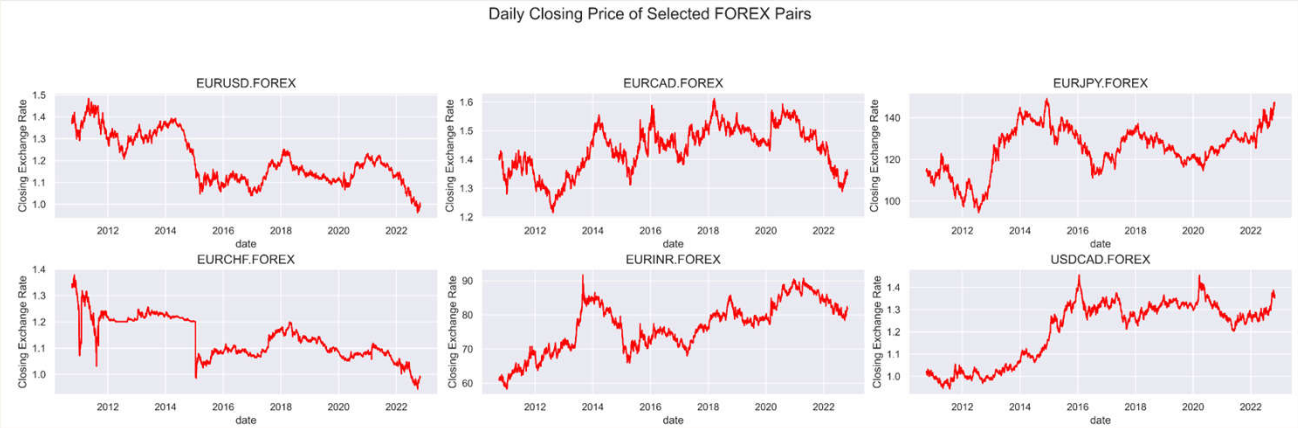
Background/Motivation

- Technical analysis (moving averages, momentum indicators, etc.) and classical indicators
- Evolving into applied ML, and deep learning (DL) as a bare minimum
- Stock price prediction a common problem, akin to demand forecasting
- DL / Reinforcement learning (RL) – “game-*ification*”
- “Side-hustle-*ification*” now enabled by many new cheap platforms (ex. Alpaca, QuantConnect)
- FOREX: interest in economic/geopolitics (implement geopolitical news trading?)
  - Less susceptible to speculation, more stable, cyclical
  - Goal: find a way to anchor forecasting to the best technical indicators – i.e., mimic the best practices of a professional trader

Workflow

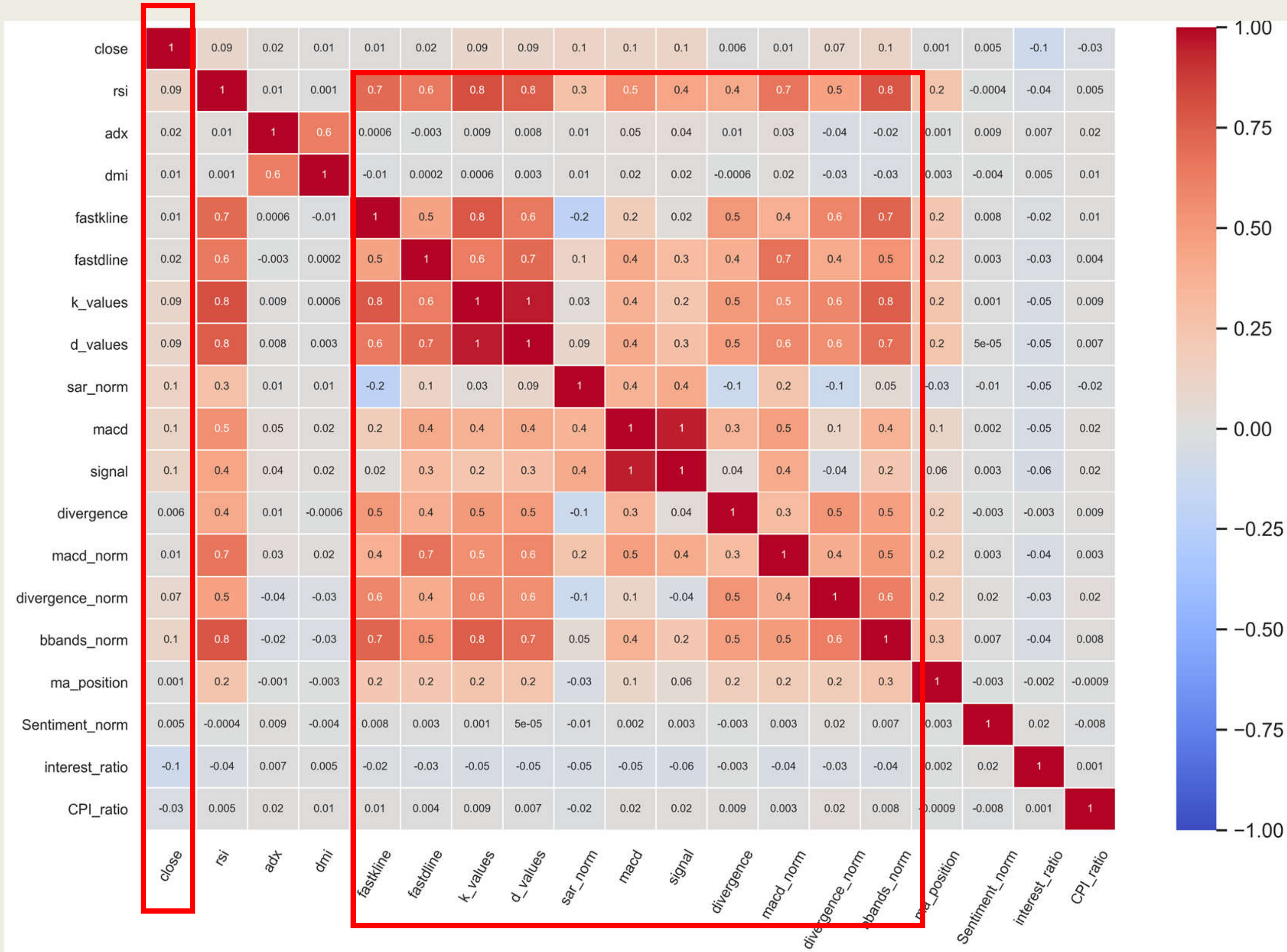
API Data + Others → Merging, Cleaning → Exploration → Classical Analysis → Prophet → Boosted Trees → LSTM

Data Gathering

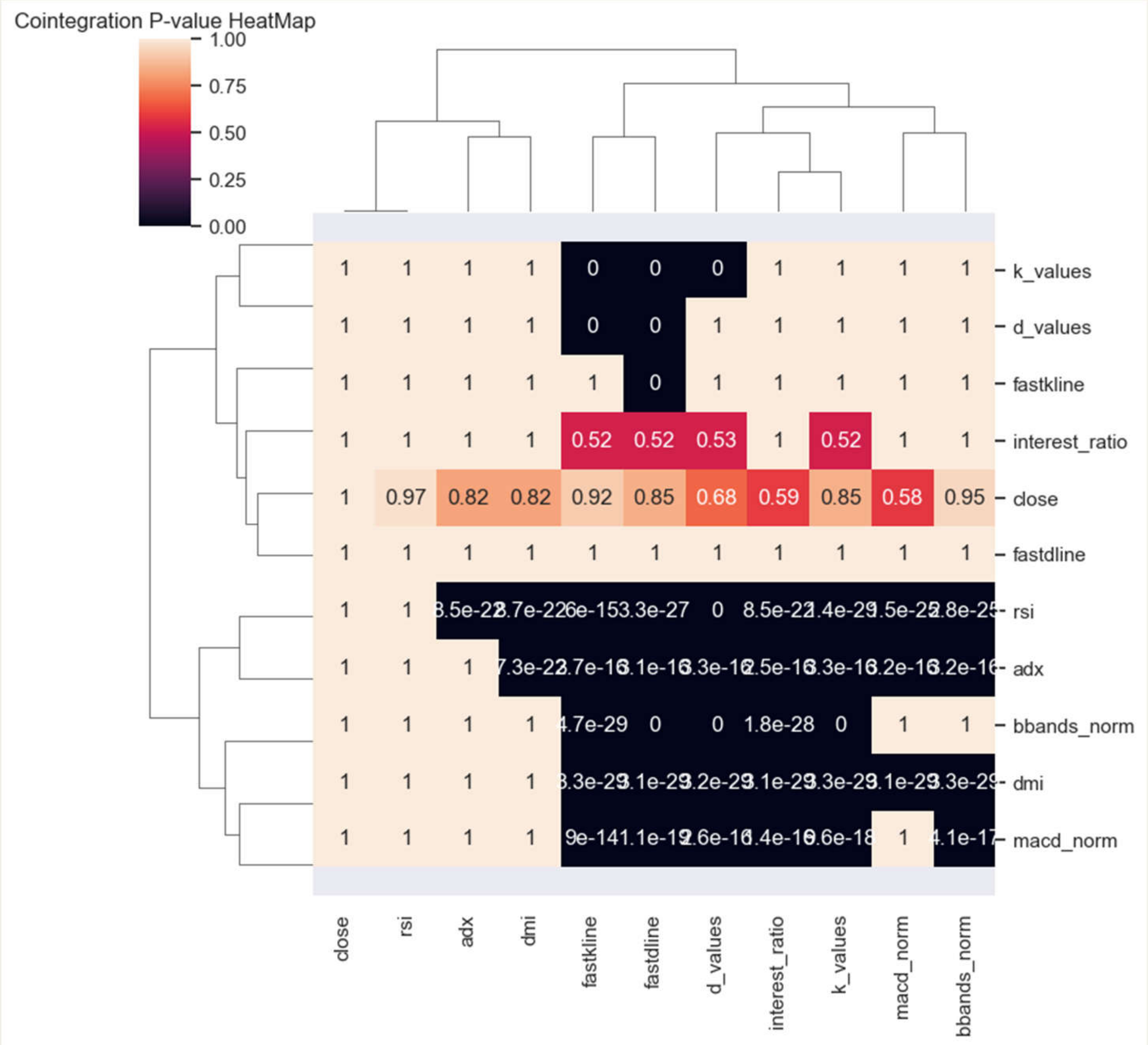


- EOD HD API
- Initially: stocks options data
- 15 FOREX pairs – 1min for 12 years: EUR – USD – JPY – INR – CAD - CHF
  - resampled daily + 15min
- EOD – Technical Indicators (RSI, moving avgs., MACD, ADX, DMI,…etc = 11)
- News Sentiment Indicator
- Interest & CPI
- Standard light cleaning and processing
  - Normalization of 5 TS
  - 9 Moving average TS into one moving average buy/sell signal TS
  - Interest & CPI ratio

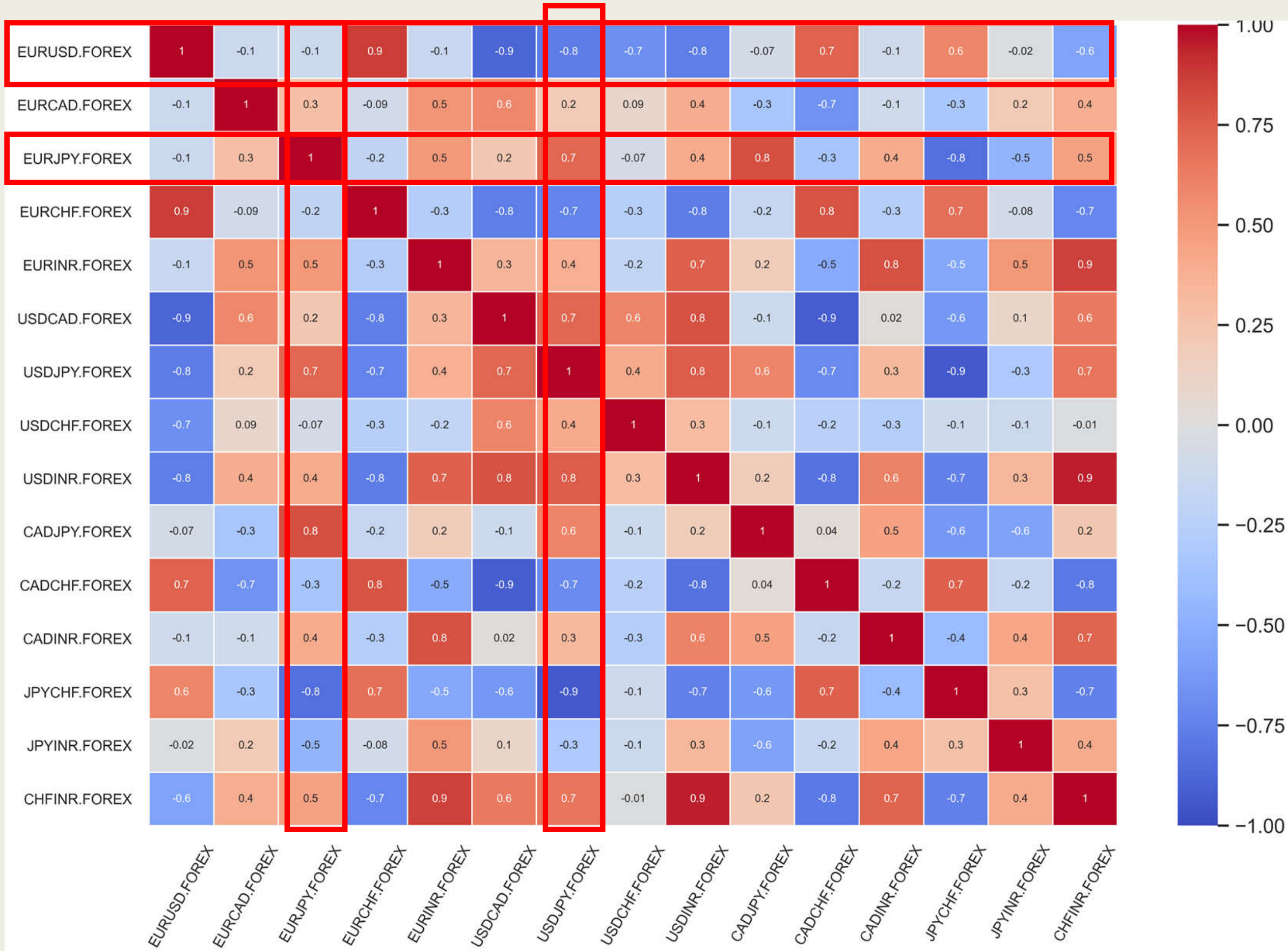
Exploration



Exploration



Exploration



Exploration

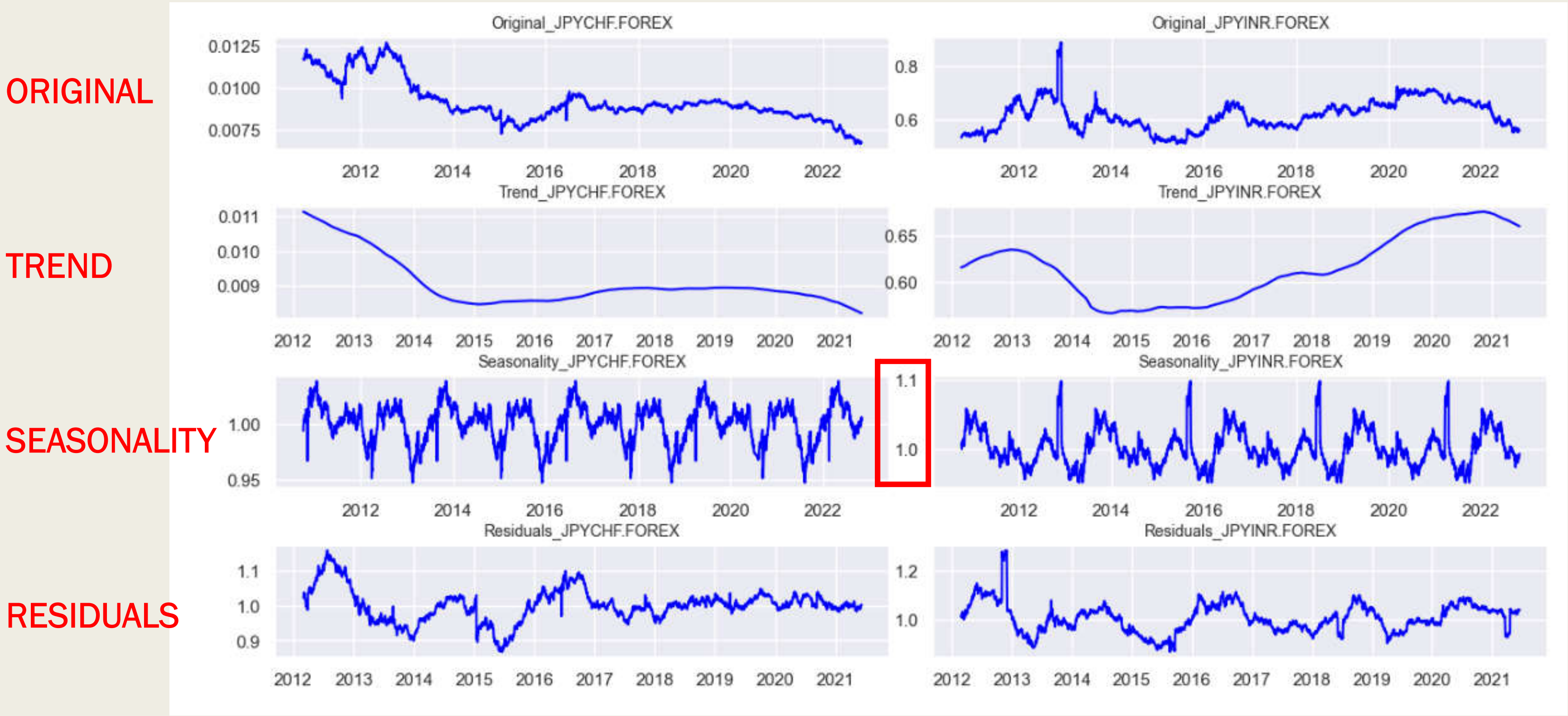
- Stationary vs. non-stationary
- KPSS statistic & AD Fuller test – recursive testing of lags for each time series
  - Lags = 7 – 365 days (smaller discretization)+ 1-3 years (larger disc.) vs. default suggested lag values in range of n=30
  - Mostly = stochastic trend only
  - = difference-stationary

Stationary?

	min p-val ADF ($p<0.05$)	lag ADF	max p-val KPSS ($p>0.05$)	lag KPSS
EURUSD.FOREX	0.101	180	0.095	638
EURCAD.FOREX	0.363	180	0.049	730
EURJPY.FOREX	0.085	270	0.100	270
EURCHF.FOREX	0.004	14	0.100	180

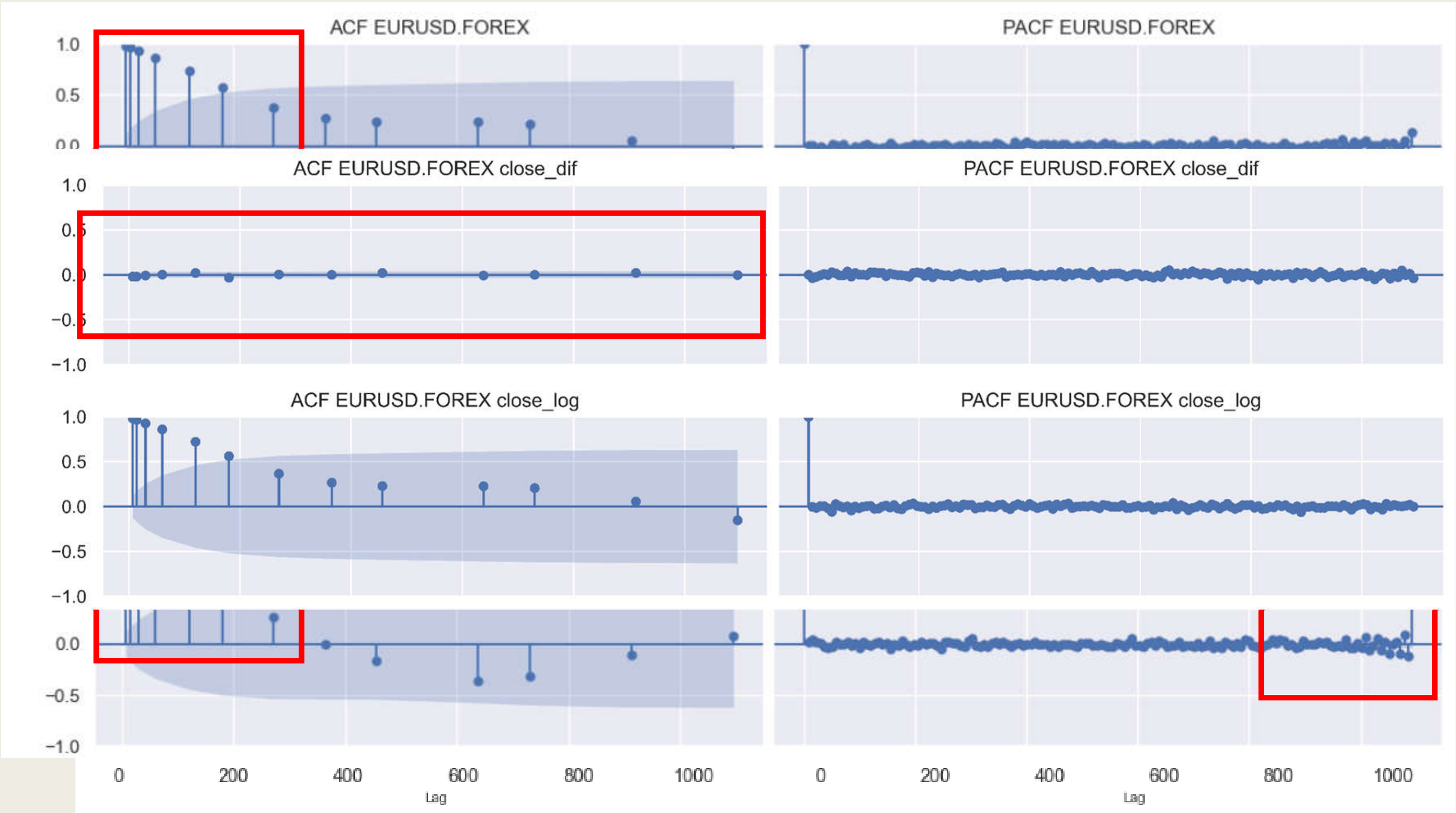
Exploration

- Long periods 1-3 years
- Multiplicative vs. additive
- +/-5-12% seasonal effect using 1-3 year periods



Exploration

- ACF/PCF



ML: VARMAX

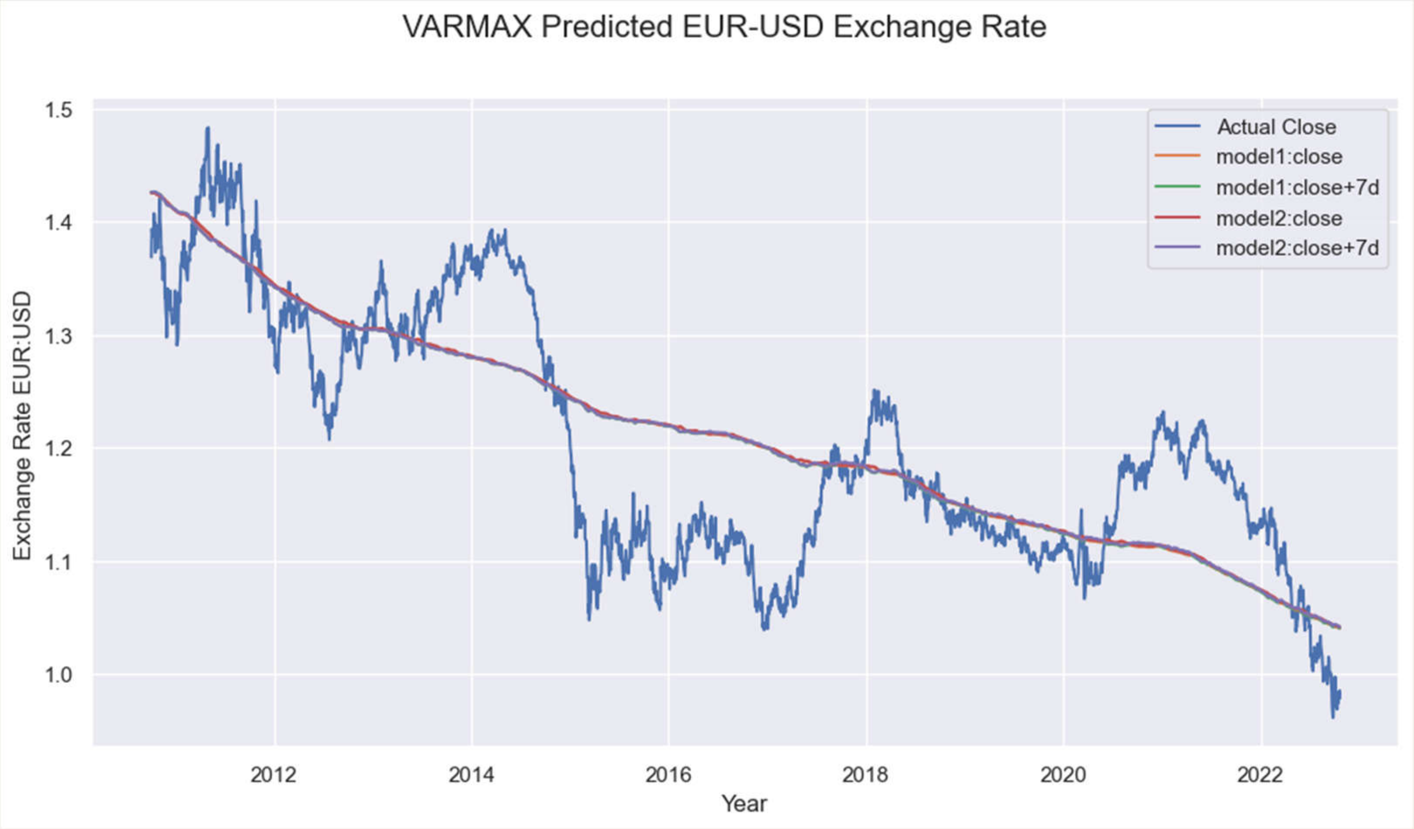
- Multivariate + exogenous variables vectorized analysis – classical model
- Requires stationarity* - but tested both differenced and undifferenced (poor performance)
- Endogenous: (close, close differenced +7d)
- Exogenous: (rsi, dmi, macd_norm, …3 to 7 total indicators) – little affect on model
- p and q = (0,3); & trend (none,constant, linear, polynomial)
- No exog var weights – removing completely had little effect

For all **models** generally used:

- Training: 2010-Oct to 2022-May
- Testing: 2022-May to 2022-Nov

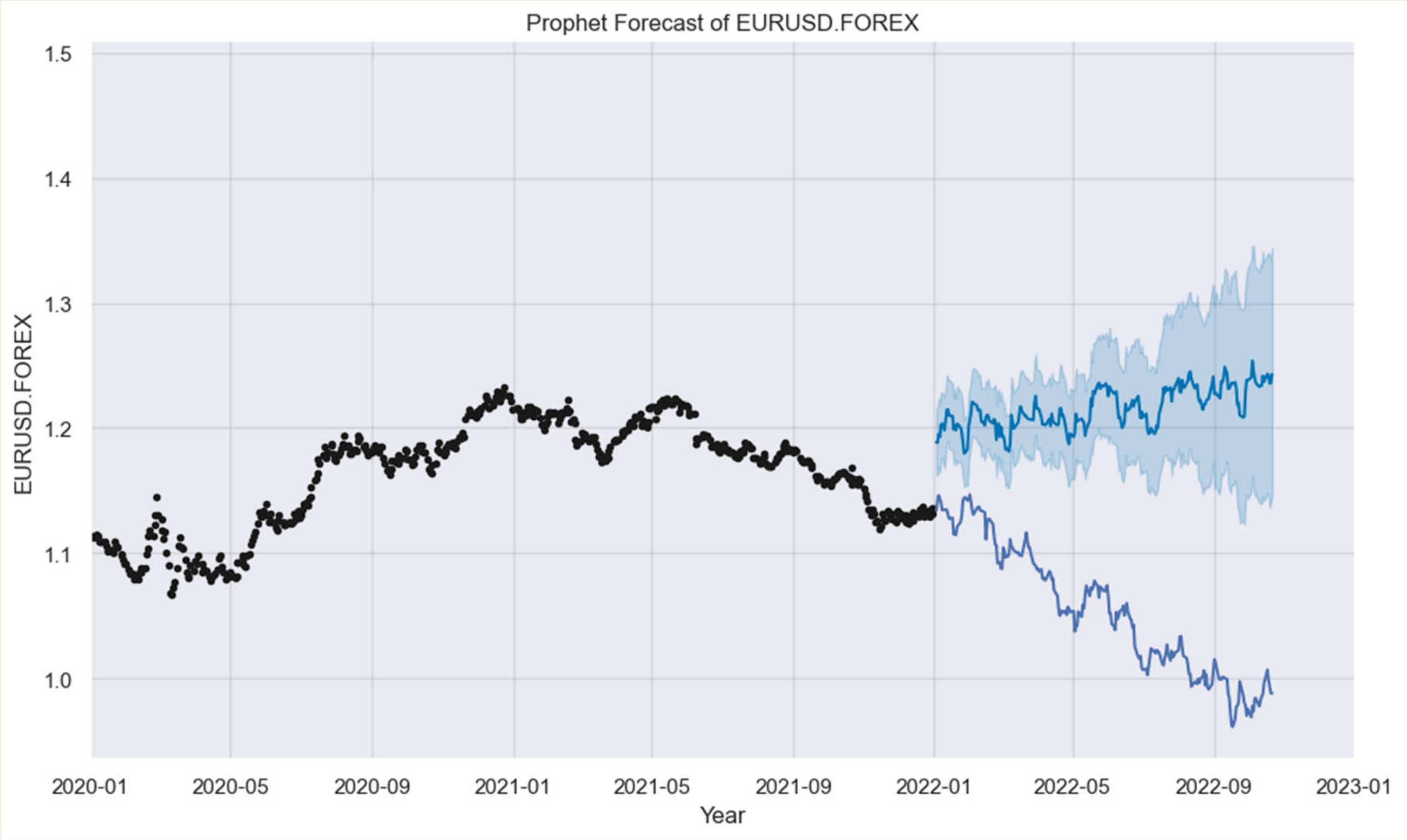
ML: VARMAX

- MAE: 0.057
- R2: 0.6



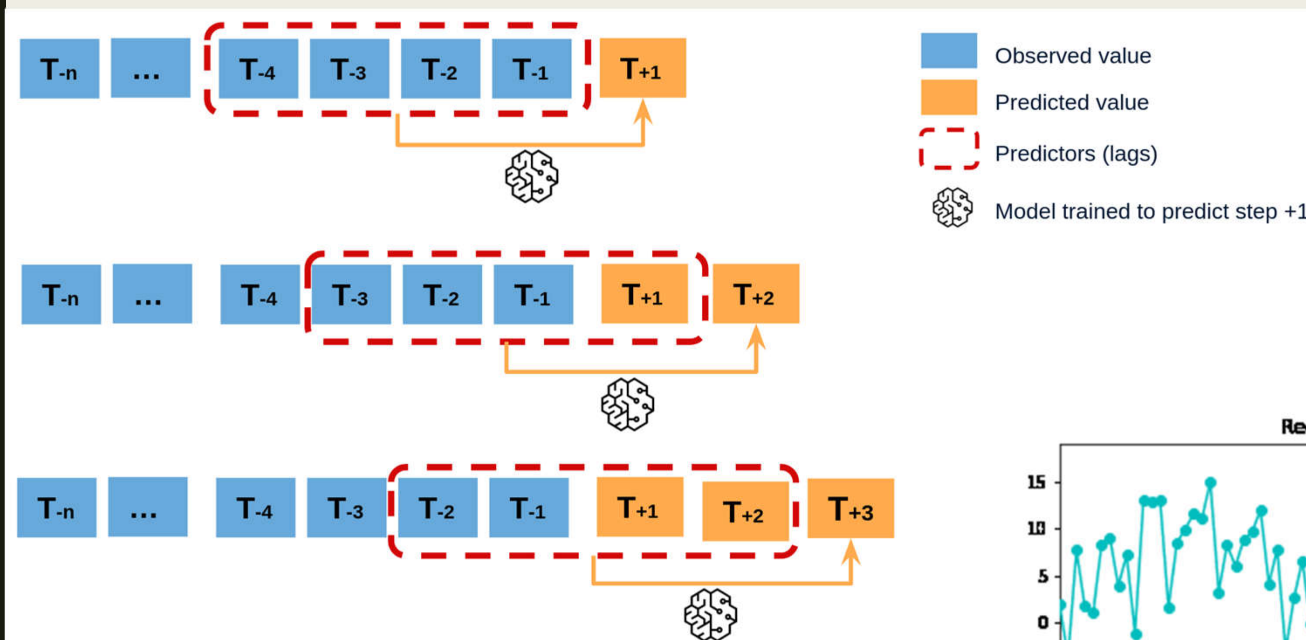
FB Prophet

- Made for business time series (demand forecasting) – strongly relies on seasonal effects

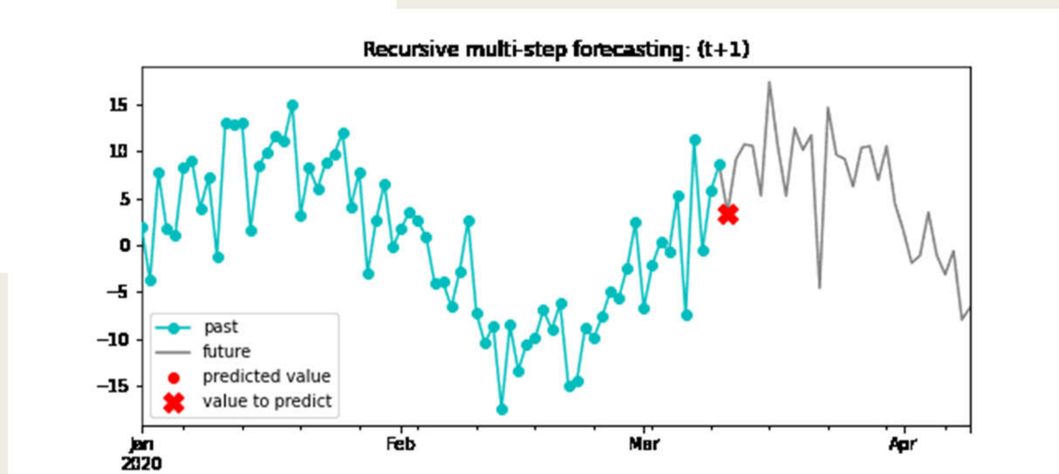


SkForecast

- Recursive multi-step forecasting – keeps relationship of lags to current step in tact
- Creates matrix for us with sklearn

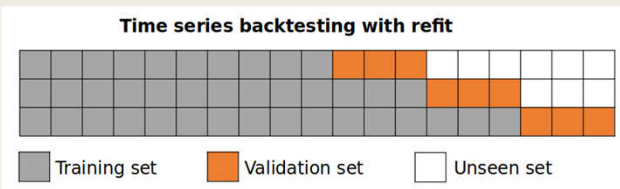


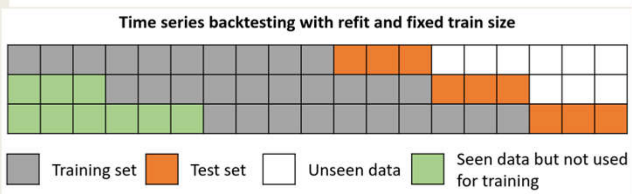
https://joaquinamatrodrigo.github.io/skforecast/0.5.1/user_guides/autoregresive-forecaster.html



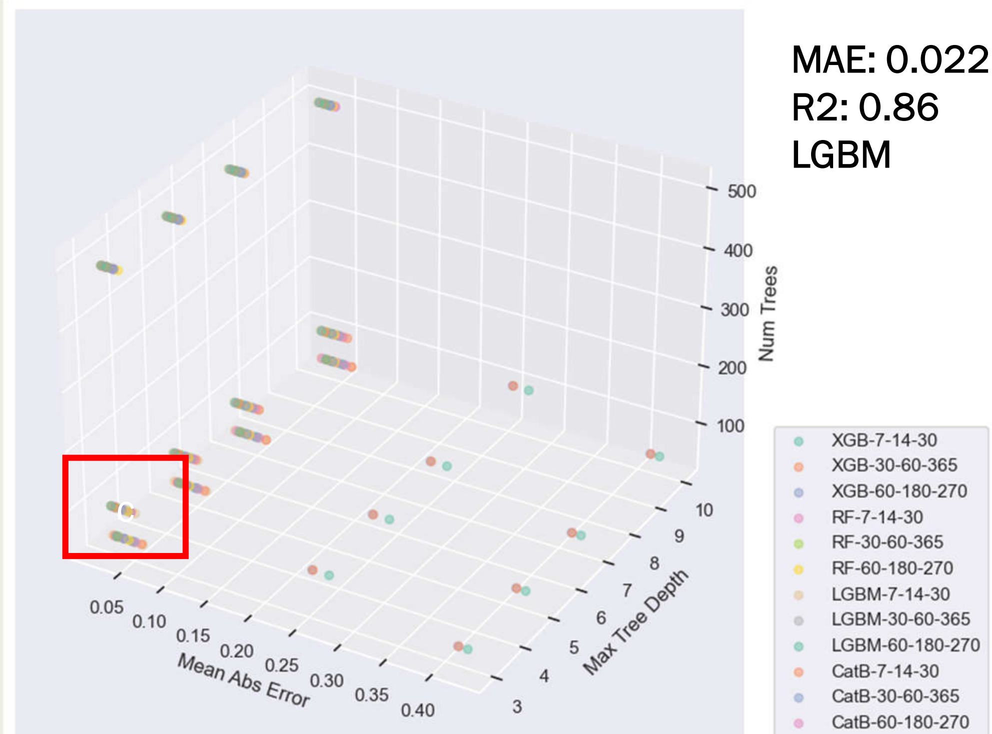
SkForecast

- Embedded Cross-validation for forecaster
- GridSearch CV hyperparameter tuning for each sklearn model
- XGBoost vs. RandomForest vs. CatBoost vs. LGBoost
- Exogenous: including all available indicators



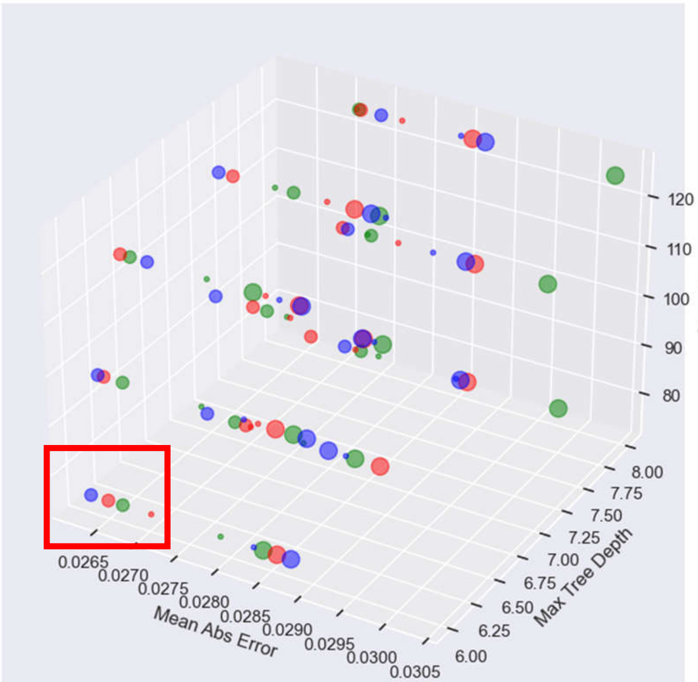


https://joaquinamatrodrigo.github.io/skforecast/0.5.1/quick-start/introduction-forecasting.html#backtesting-with-refit-and-fixed-training-size-rolling-origin

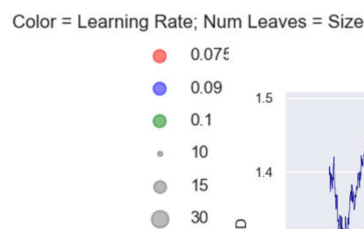


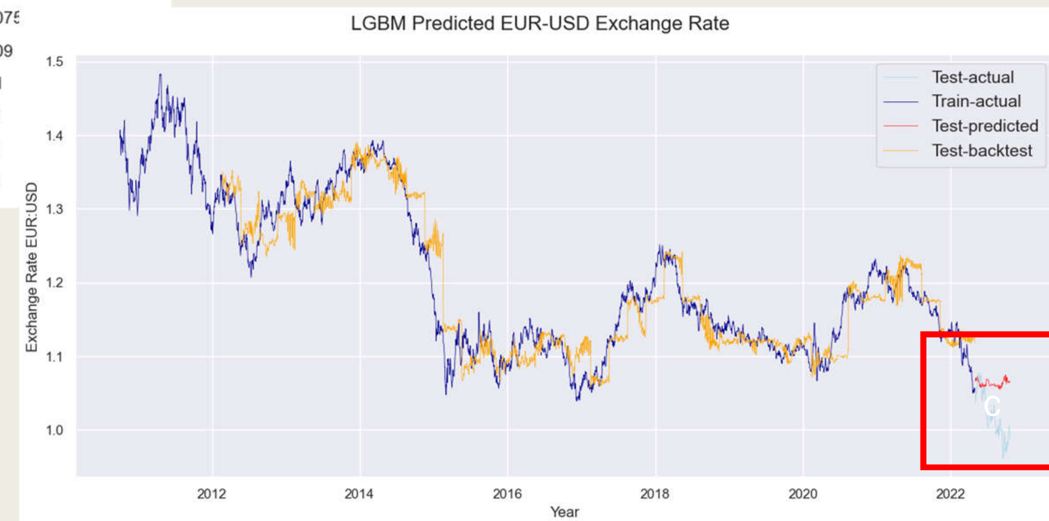
MAE: 0.022
R2: 0.86
LGBM

SkForecast



- LGBM only
  - n_leaves
  - Tree depth
  - Learning Rate
  - Estimators (n Trees)

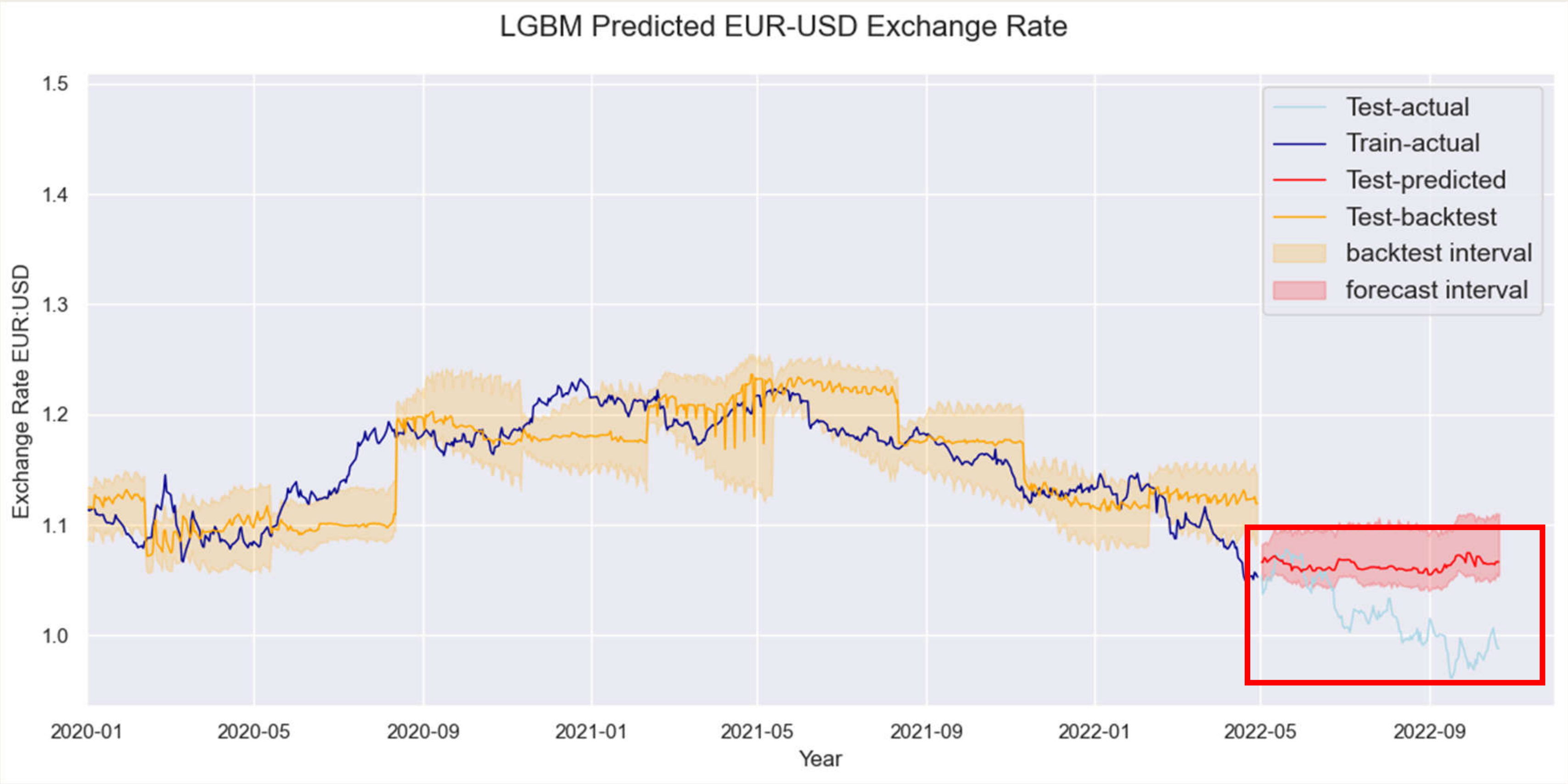




Backtest
MAE = 0.024
R2 = 0.87

Forecast
MAE = 0.046
R2 = - val ☹

SkForecast



Lags in model = 95% of feature importance

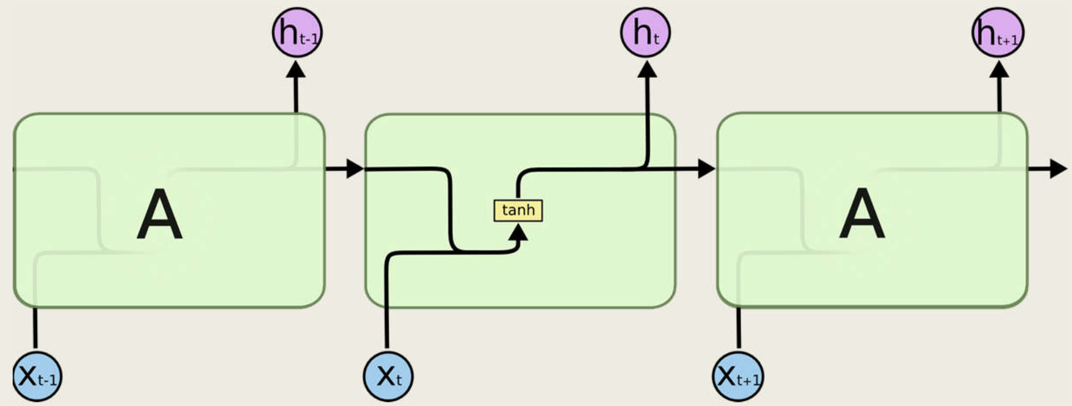
LSTM - Keras

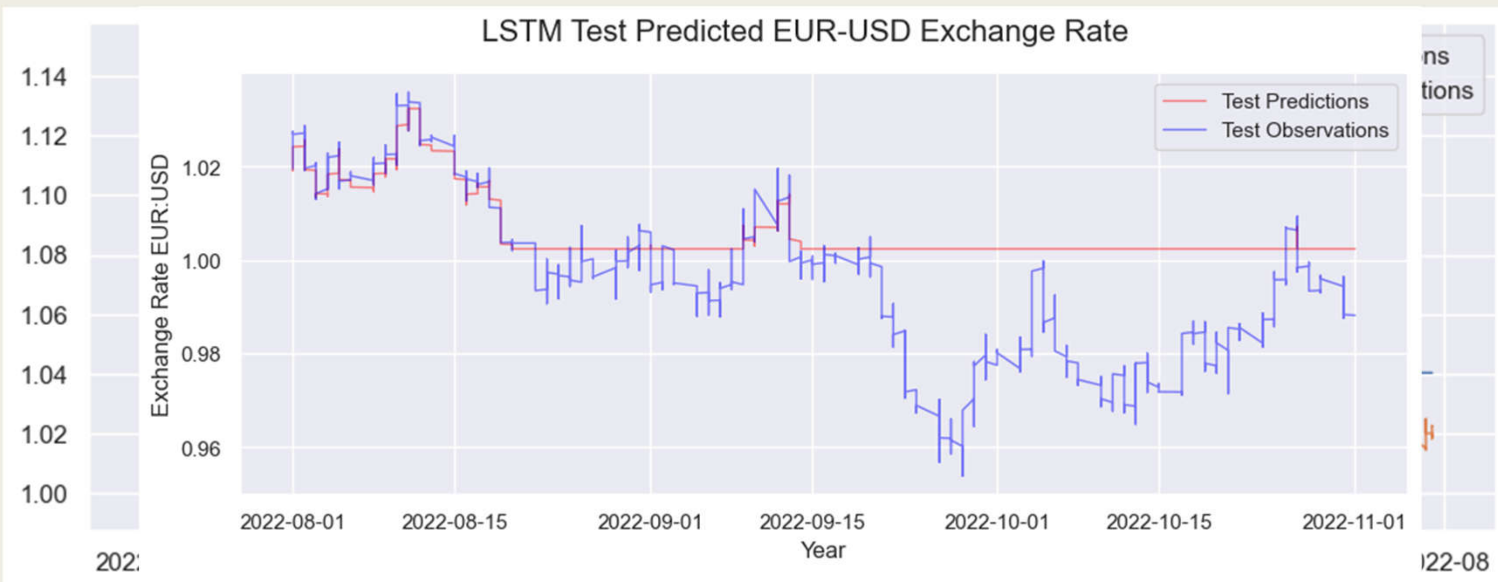
- RNN
- Exog = [7,14,30]d lags

Exponential Moving Averages [50, 100,200]d lags

MAE: 0.012

R2: 0.14





Conclusion

- Tested wide array of forecasting models'
- Technical indicators not absorbed by models

- Combination of LSTM: short term price prediction
- Try to refine boosted trees: Run several exchange pairs – exog. and multi-var endog.
- LSTM + other models
- Voting
- Deep Reinforcement Learning