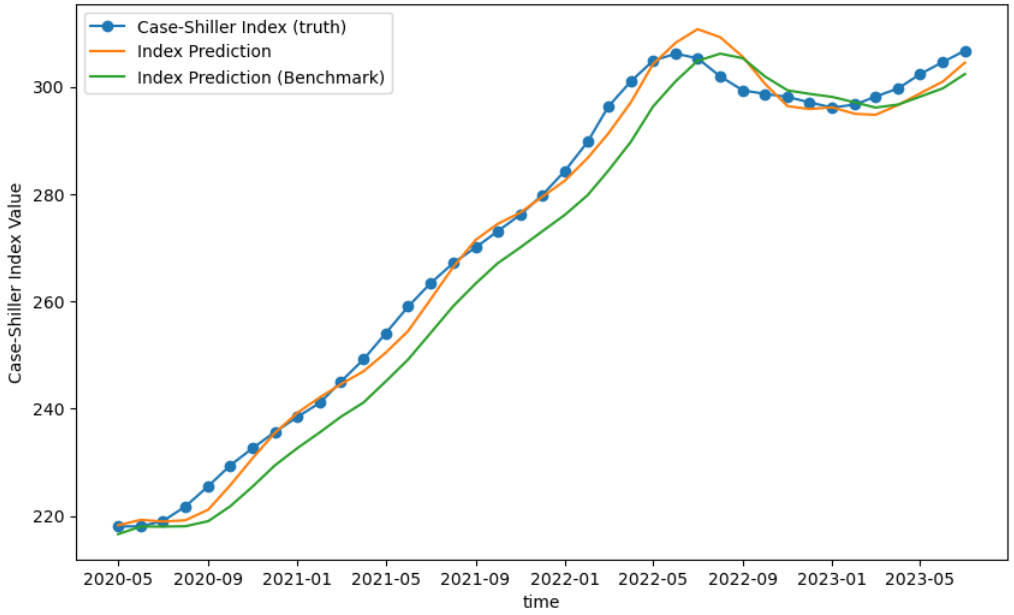
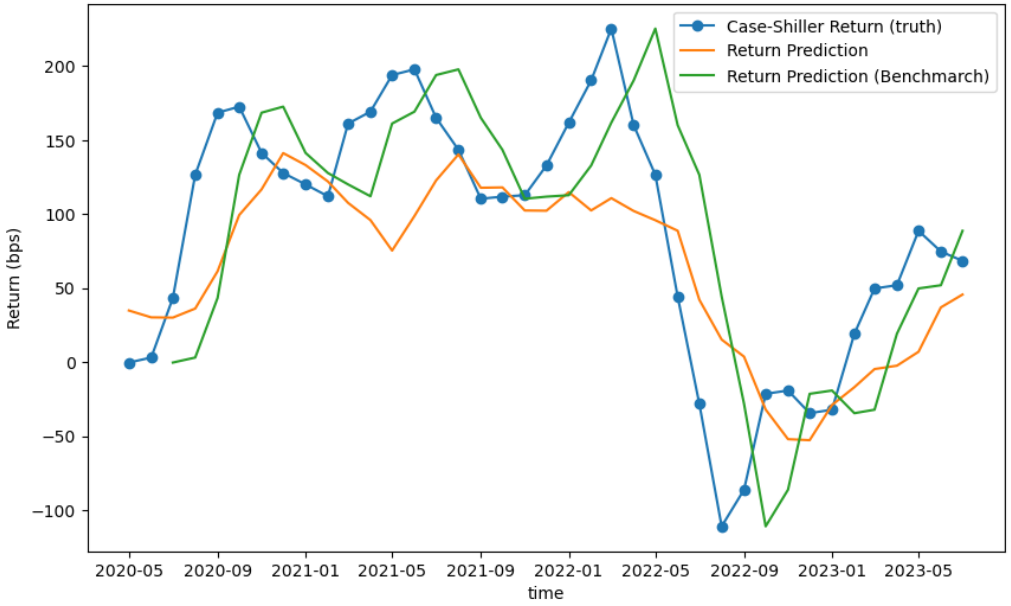
S&P Case-Shiller Index Prediction

Siyun

Summary

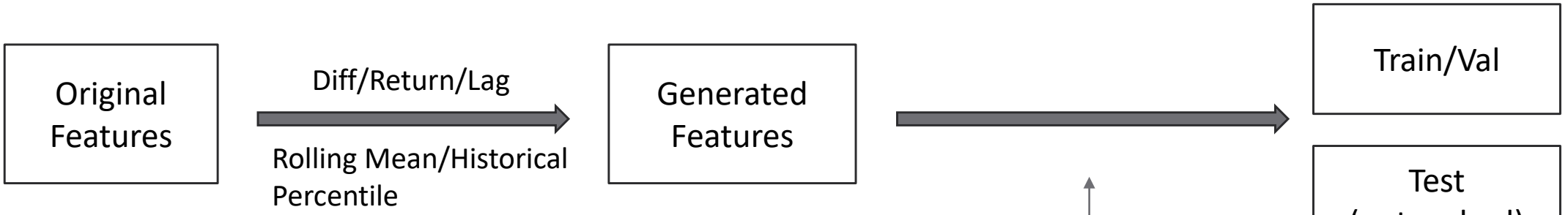
Case-Shiller Index is usually released with ~2 month lag.
We aim to predict its accurate value at current month (also known as now-casting)





We use LightGBM to train a point-in-time dataset with 1000+ features from 6 categories.
Our prediction is better (test RMSE: 8.77), and more responsive than the benchmark (test RMSE: 39.86) which uses latest available index value as prediction.

Features Engineering



44 original features

- 6 Categories: Monetary/Economic/Demographics /Employment/Housing/Markets

1081 derived features

- Captures trend, relative position, etc

"Time in Point" Dataset Construction

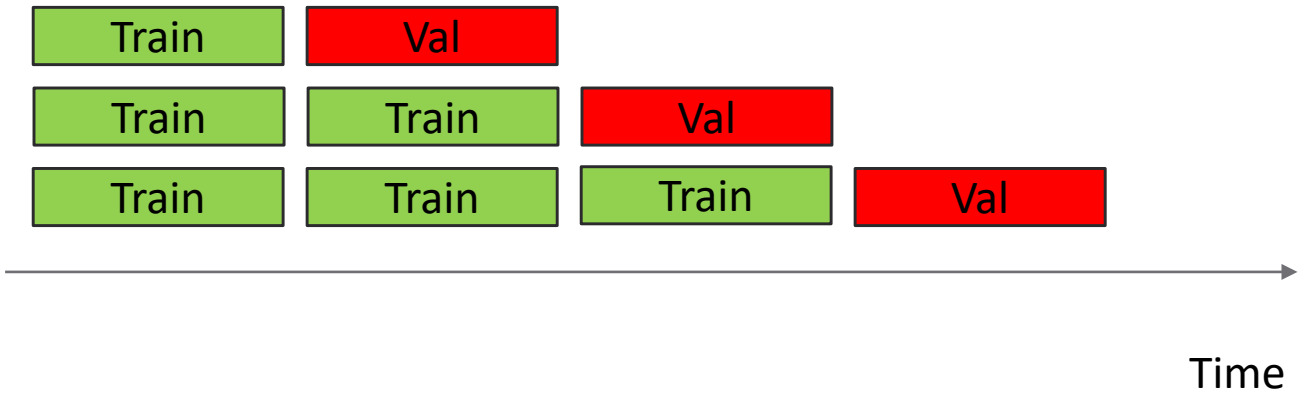
- Each row in data will have latest information up till the corresponding month
- Ensures no look-ahead bias
- We process ragged-edge features using **"vertical alignment"** method (research shows similar performance vs using predicted value)

Train Methodology

Model: LightGBM

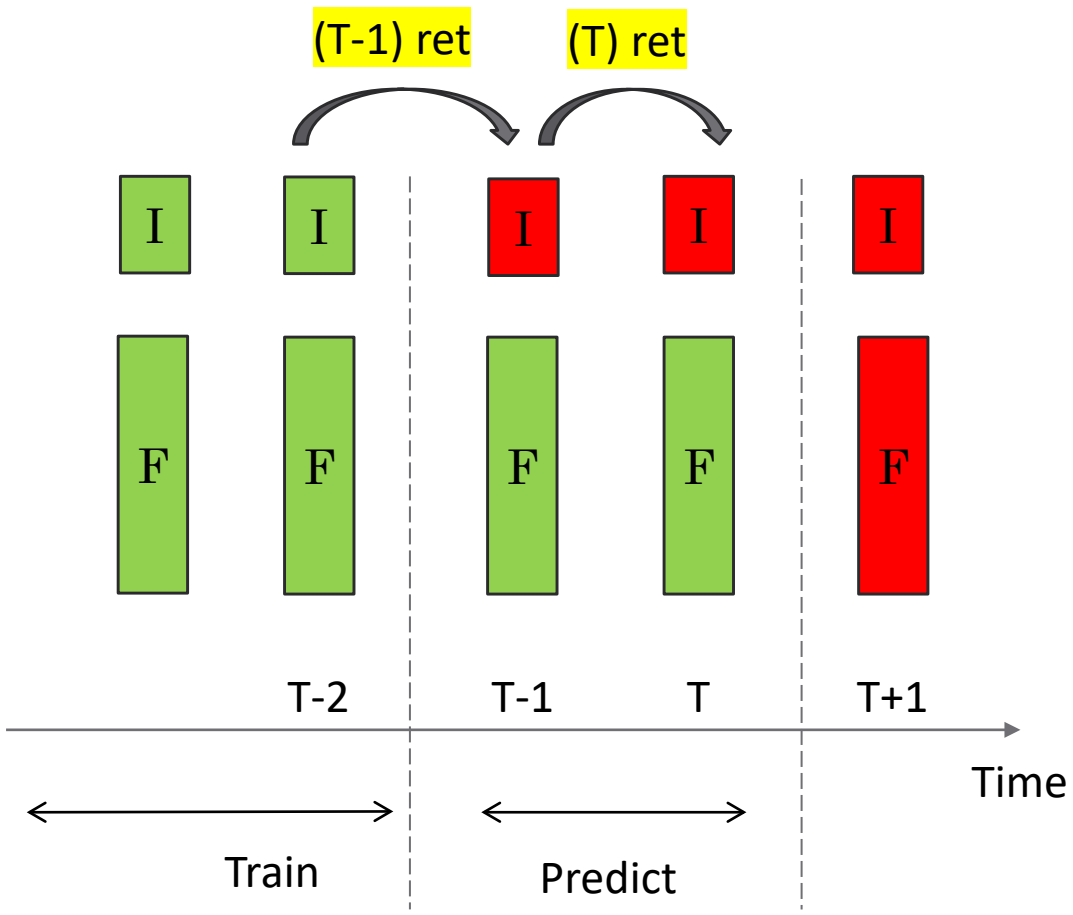
- Relevant research shows commonly used methods for economic data (now/fore)-casting include:
  - Dynamic Factor Model / Timeseries / ML
  - ML, especially boosting methods, are usually proven to outperform.
- **LightGBM** is chosen due to its performance, efficiency and null data handling capability
- Prediction target is set to be **index return** rather than index value (non-stationary).

Training Scheme: Time Series CV & Params Tuning



- 5-Fold Time Series CV is conducted
- Hyper-param is tuned based on RMSE loss, using Bayesian scheme (optuna)

Prediction Methodology : Expanding Window



F Feature set, known

I Index, known

F Feature set, unknown

I Index, unknown

Currently at Month T

- To predict the actual value of the index, we look at the latest available Case-Shiller index data available at the given date of T month
- If latest available data is (T-2) month, then:

predicted index at (T) month = (T-2) index * (T-1) ret * (T) ret

Then, move to next Month T+1

- Training window expands 1 month to include T-1
- Retrain model with newest data
- Predict window moves 1 month to be (T, T+1)