Project Work in Machine Learning: M5 Competition

Michele Vece

Summary

01 **M5 Competition Overview**
Objective, organization and metric

02 **Main challenges**
Issues and solutions

03 **Other solutions**
Lessons learned from other participants

04 **Proposed solution**
Description of the solution

1.1: M5 Competition

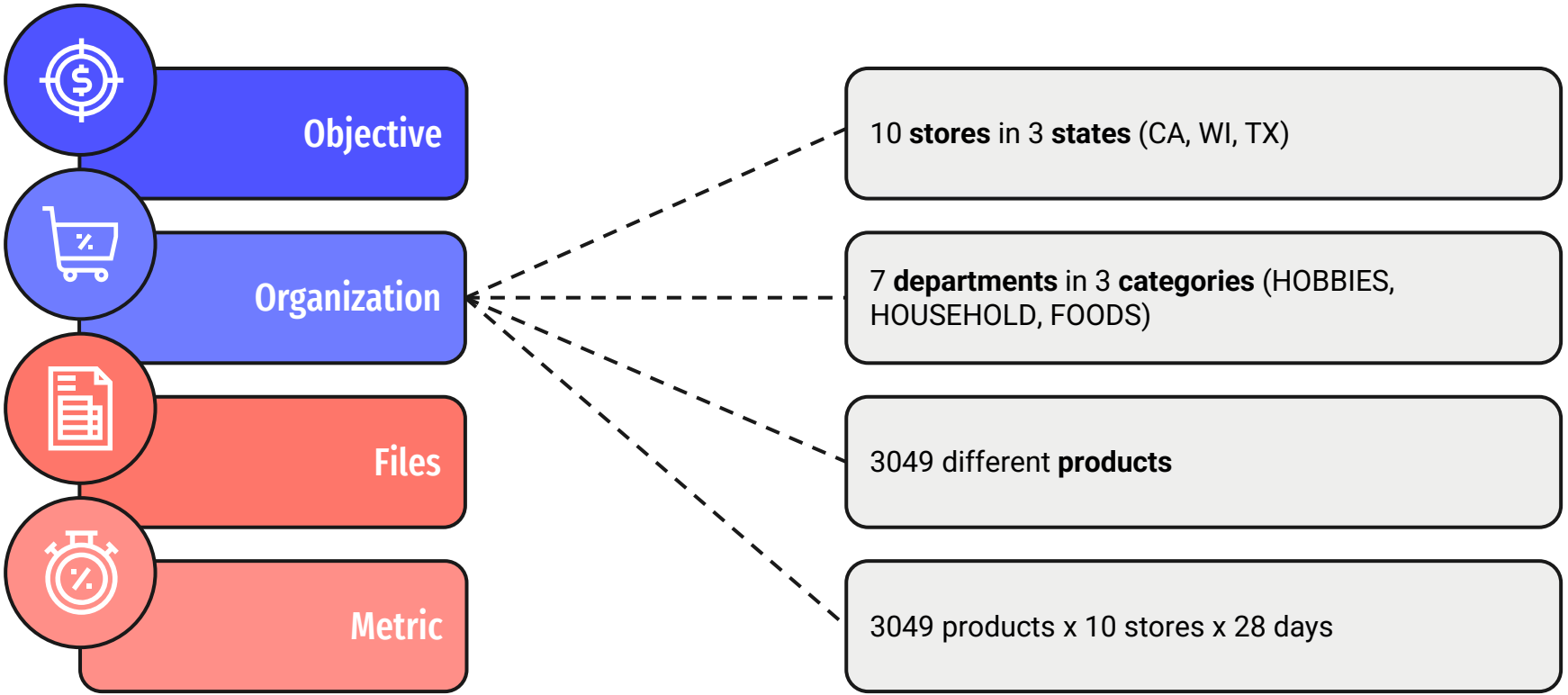
Unit sales forecasting:
for each item in each store for the next 28 days

Historical sales:
days 1 - 1941

Public leaderboard:
days 1914 - 1941

Private leaderboard:
days 1942 - 1969

1.2: M5 Competition



1.3: M5 Competition

calendar:
dates the products are sold

sell price:
weekly price of products in each store

sales train validation:
historical daily unit sales (until day 1914)

sales train evaluation:
historical daily unit sales (until day 1941)

1.4: M5 Competition

WRMSSE:
Weighted Root Mean Scaled Squared Error

RMSE scaled wrt **historical sales**

$$RMSSE = \sqrt{\frac{1}{h}\frac{\sum_{t=n+1}^{n+h}\left(Y_t - \hat{Y}_t\right)^2}{\frac{1}{n-1}\sum_{t=2}^{n}\left(Y_t - Y_{t-1}\right)^2}}$$

RMSSE weighted wrt **cumulative actual dollar sales**

$$WRMSSE = \sum_{i=1}^{42,840} w_i * RMSSE$$

12 aggregation leves:
Overall score is the average of 12 WRMSSEs

2: Main challenges

01

Amount of data

Feature Engineering:
- high-level features on aggregated data
- divide in subsets
- low-level features on each subset.

Model implementation:
- divide in subsets
- train a different model on each subset

Here:
- Split data wrt state
- FE
- Split date wrt store
- training

02

Cross correlation

Similarities among different timeseries:
- same store, different department
- same department, different store

How to split data?
- subsets that share some common behaviour may be separated

03

Sales intermittency

Are zero sales real?
- out-of-stock

Solution n. 1:
- predict sales
- predict out-of-stock probability

Solution n. 2:
- objective function that works well with non-negative right-skewed distributions

04

Prediction Atomicity

Atomicity:
- product-level
- daily basis

Error-prone:
- easier to make errors than on aggregated data

Overall score:
- takes into account errors at different levels of aggregation

05

Horizon, Recursion

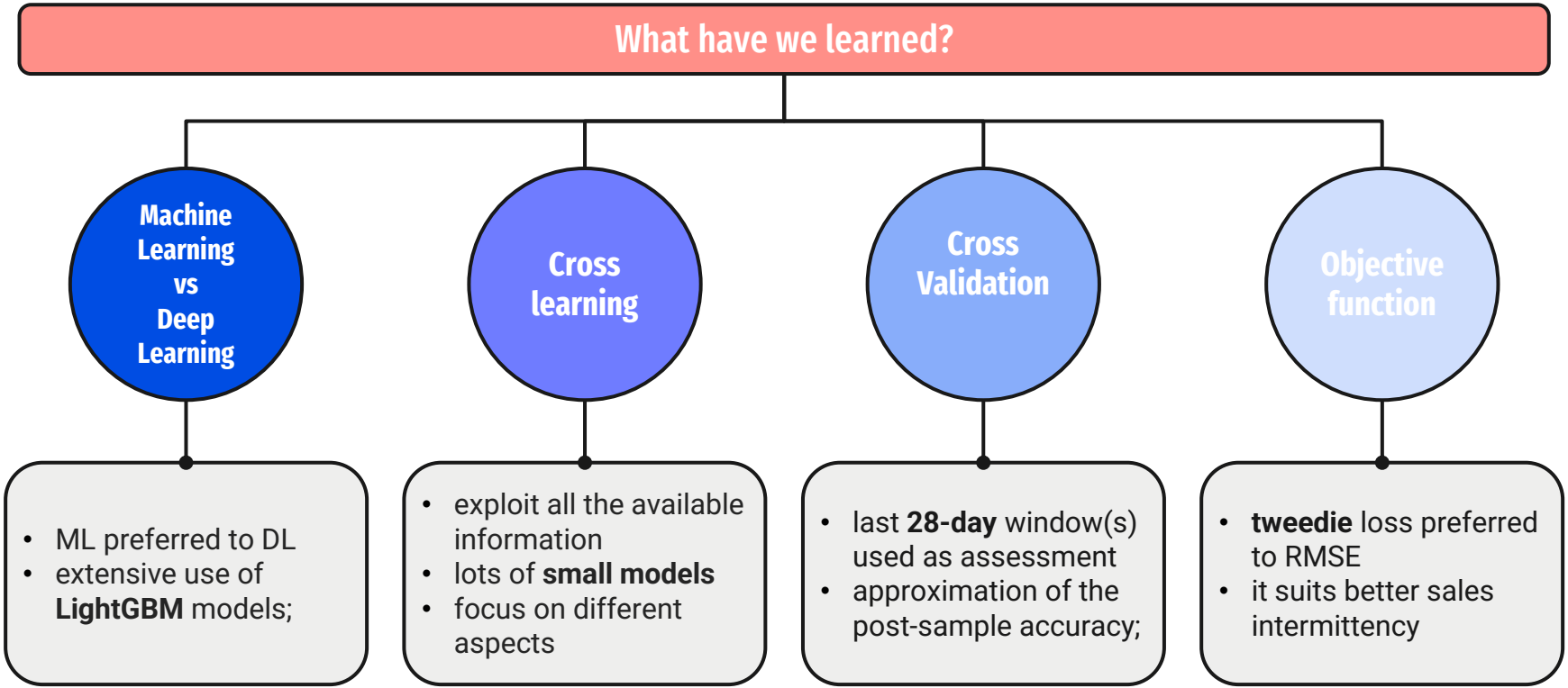
General rule:
- latest days have higher predictive power

Recursive data:
- fresh data
- may contain error

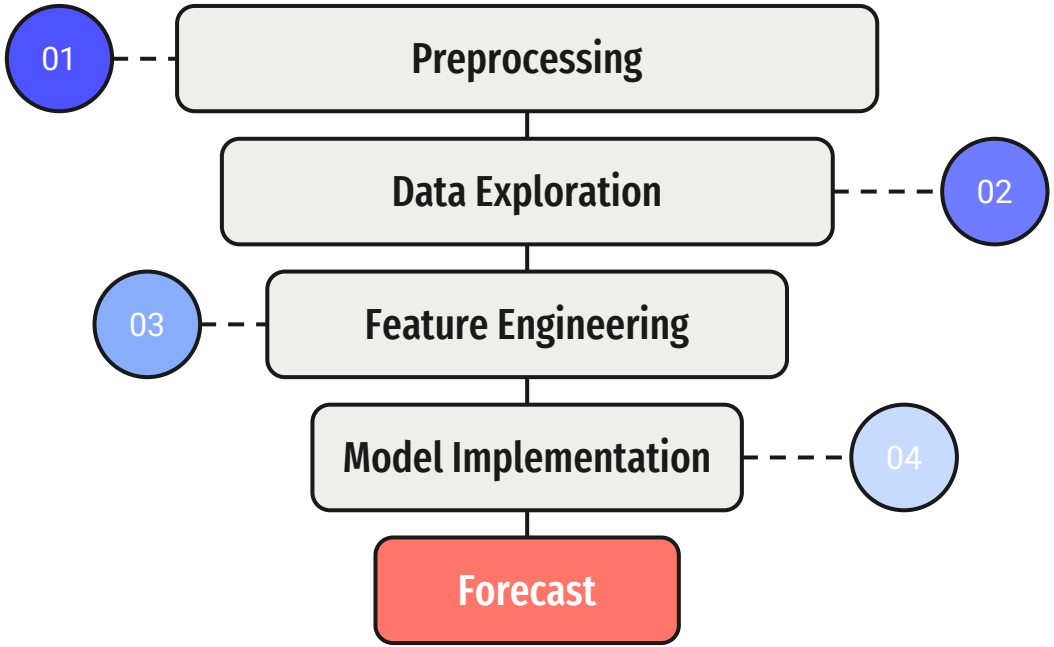
Non recursive data:
- less recent
- ground truth

3 models:
- a model per *all* weeks
- a model per *each* week
- recursive model

3: Other solutions



4: Proposed solution



4.1: Preprocessing

Calendar

Dates 1 - 1969

- date
- wm_yr_wk
- weekofmonth
- weekofyear
- dayofweek
- dayofmonth
- dayofyear
- month, year
- event_name_1
- event_type_1
- snap_CA
- snap_TX
- snap_WI

Prices

Prices for each item, store

- store_id
- item_id
- wm_yr_wk
- sell_price

Sales

Historical unit sales

- id
- item_id
- dept_id
- cat_id
- store_id
- state_id
- d_1
- ...
- d_1941

4.2: Data Exploration

Sales distribution

Unit sales grouped per state, store, category, department
- different distributions
- not uniform

Cross correlation

Similarities across same store or department
- group/divide per store
- group/divide per department

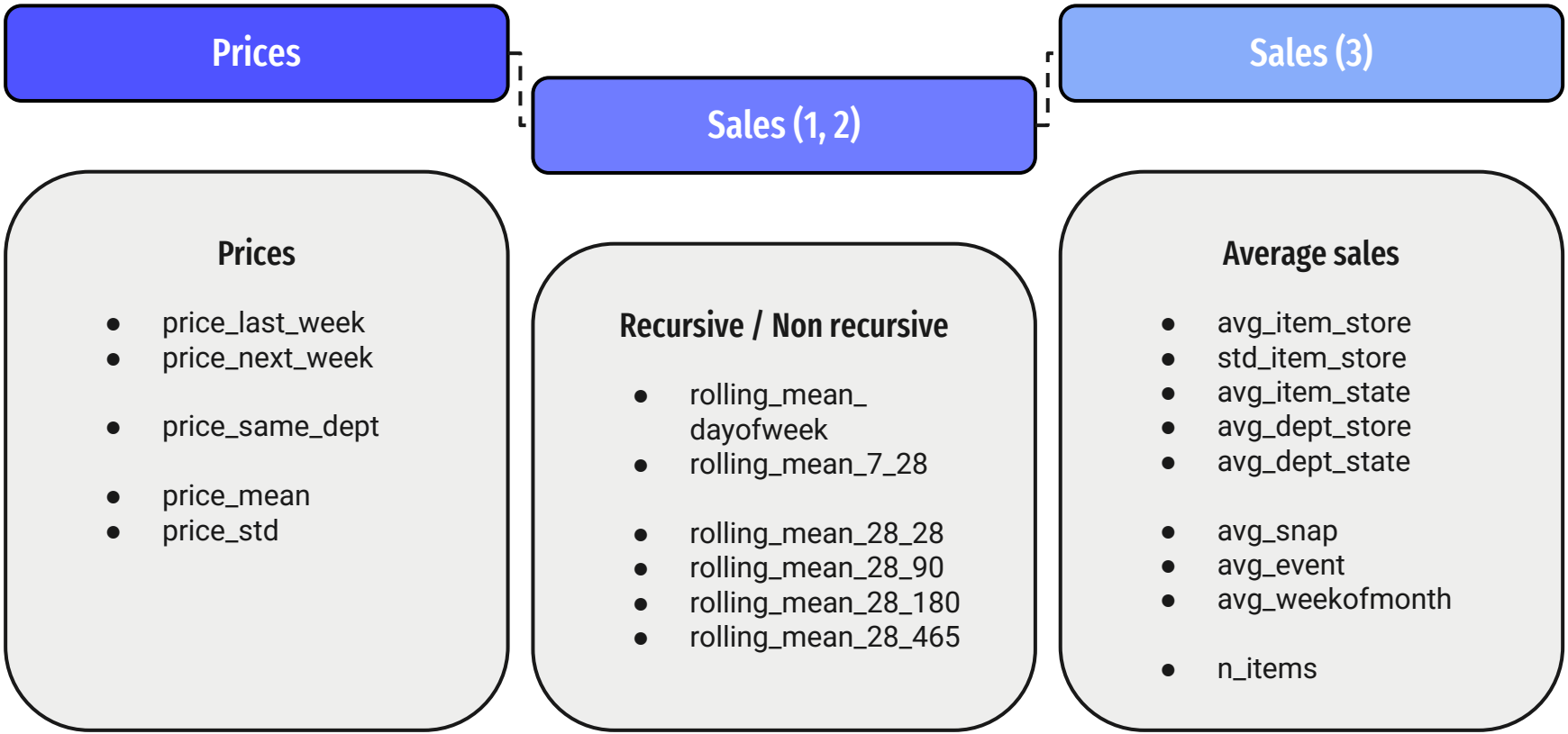
Seasonality

Unit sales per period and event
- higher sales on weekends, on snap days, on some events, …

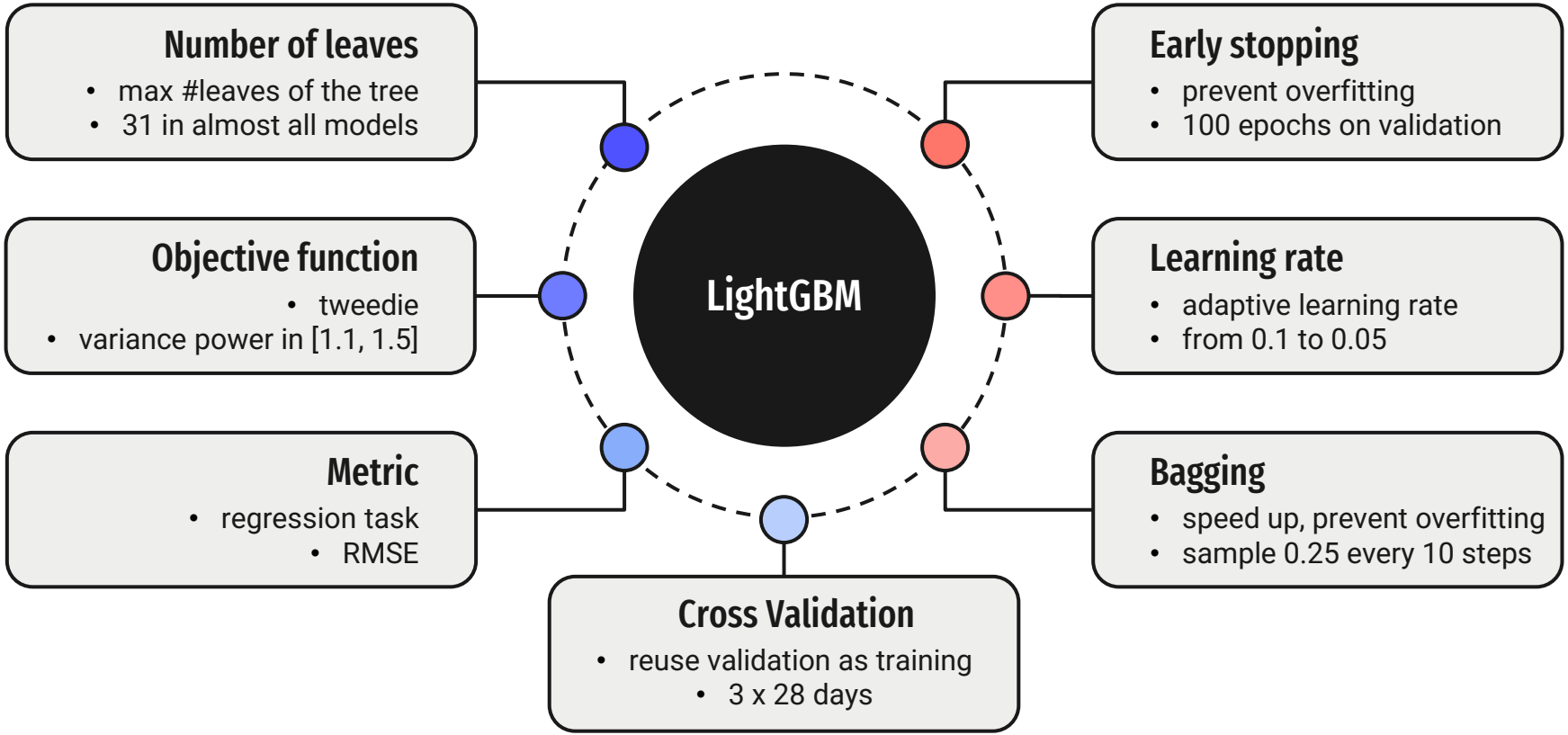
Autocorrelation

Discover useful lag values
- lag 7 and its multiples
- lag 30/31

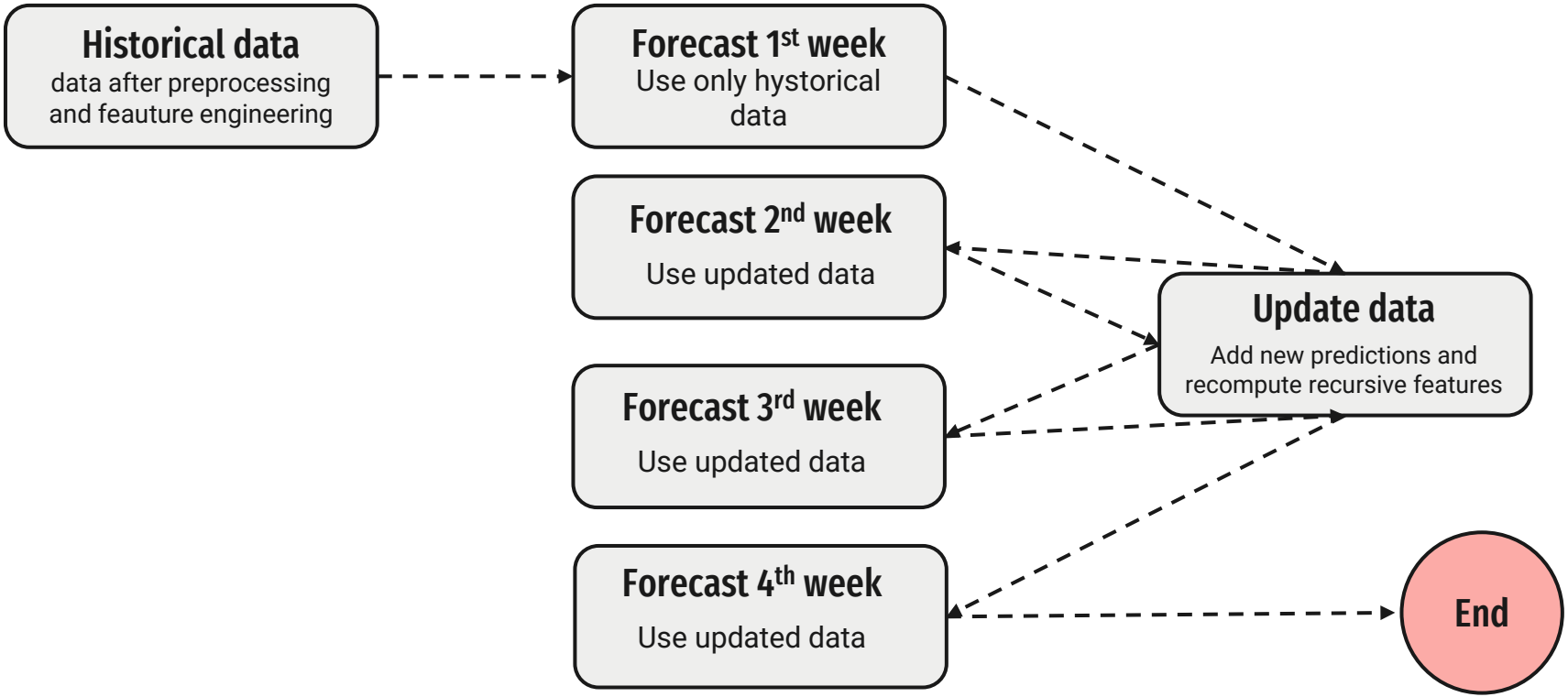
4.3: Feature engineering



4.4: Model implementation



4.5: How recursive predictions work



4.6: Submission and results

YOUR RECENT SUBMISSION

m5_final_submission.csv
Submitted by Michele V · Submitted 19 hours ago

Score: 0.57415
Public score: 0.60636

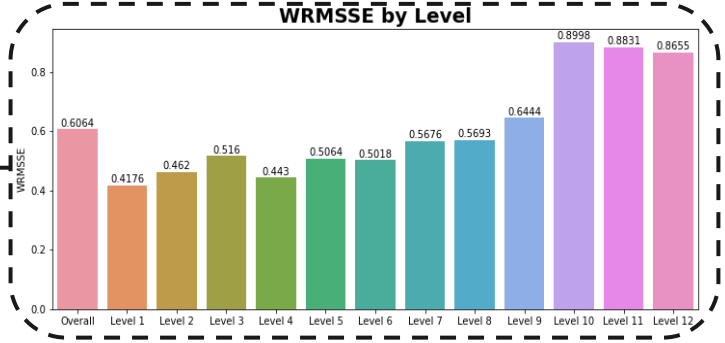
Jump to your leaderboard position

Leaderboard

- top performing benchmark beated
- similar private/public

WRMSSE decomposition

- deterioration at lower aggregation level



45	▲ 1665	Hiromitsu Kigure		0.57378
46	▲ 360	YK		0.57457

Placement

- 46th place in the public leaderboard