Recap

25 February 2025 17:07

Demand Prediction → # of pickups

Dataset → NYC Taxi → 2016 → Jan, Feb, March } 2GB

Row → single ride → fare, distance, Pickup coord, Drop coord

Region, Time → System → # of pickups

Outliers → Remove

Dask → chunks → operations

↓

Removed the outliers.

✓ Pickup datetime — time, date

lat pickup

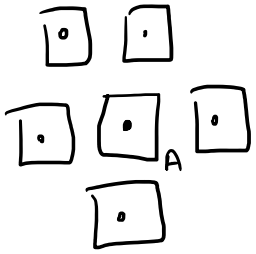
long pickup

3·3 crore

Task 1 → Break our NYC into regions

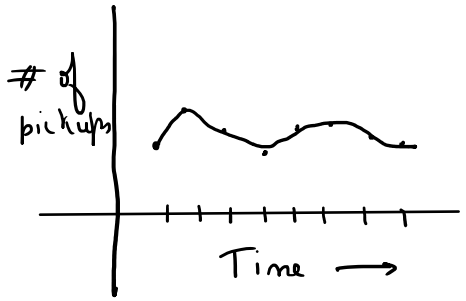
unsupervised ML → clustering

Task 2 → Historical data for pickups for each region.



Time interval — what are the # of pickups

Time axis → 15 min intervals



3.3 crore lat long

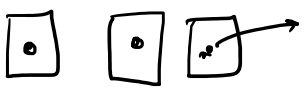
Task 1 → Mini Batch K Means → Batch size

Batch → Centroid distance

Pandas chunking → Partial fit

chunk → Batch → centroids

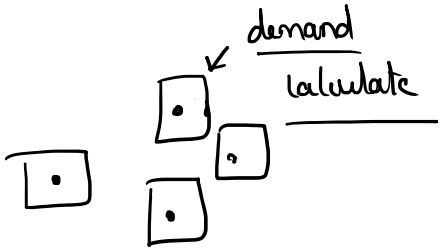
Regions → Centroid coord (Region centers)



Task 2 → Region Datetime

15 min interval → Resample

15 min interval → Resample



calculate for a given 15 min interval.

8:00pm – 8:15 pm.

extra feature → avg demand avg-pickups

for each region and for each 15 min interval.

EWMA → Historical data + Current observation O_t

↓ Weight ↓ Weight

→ exponentially decaying in nature.

Datetime	Region	total pickups	avg-pickups

derived feature.

Flow

25 February 2025 17:08

datetime	Region	avg pickups	total pickups (target)
—	—	—	—
—	—	—	—
—	—	—	—
—	—	—	—
—	—	—	—

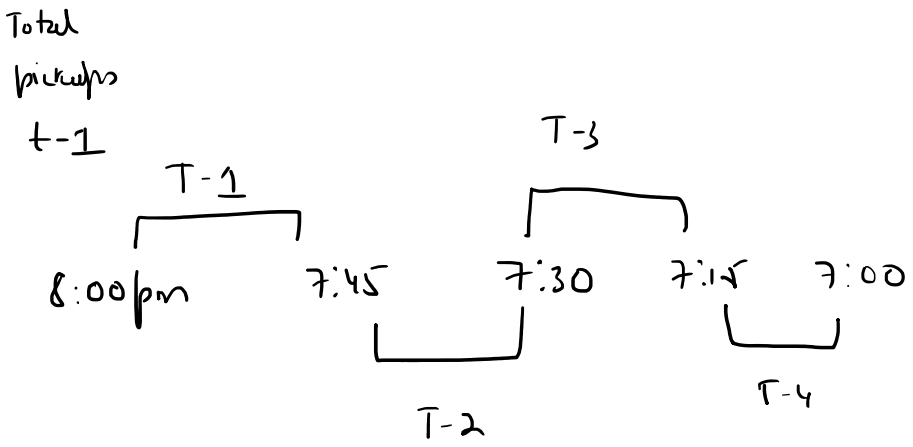
✓ Region

✓ Time interval

Features → lagged features.

T_0 → Current time interval.

T-1 T-2 T-3 T-4



input data T-1 T-2 T-3 T-4 Region Avg Pickups

train test split

time axis

Month → Jan, Feb (Train set)

March (Test set)

→ For every region and every 15 minute interval.

EWMA $\alpha = 0.9 \rightarrow 0.4$

Last N timesteps EWMA $\alpha = \frac{2}{N+1}$

last 4 timesteps.

$t-1, t-2, t-3, t-4$ $\alpha = \frac{2}{4+1} = \frac{2}{5} = 0.4$

EWMA = 0.4

We are actually looking into historical data

EWMA Avg → Some part of history in the average

10 → EWMA features

Data leakage

$\alpha = 0.4$ → Historical avg of prev 4 timesteps.