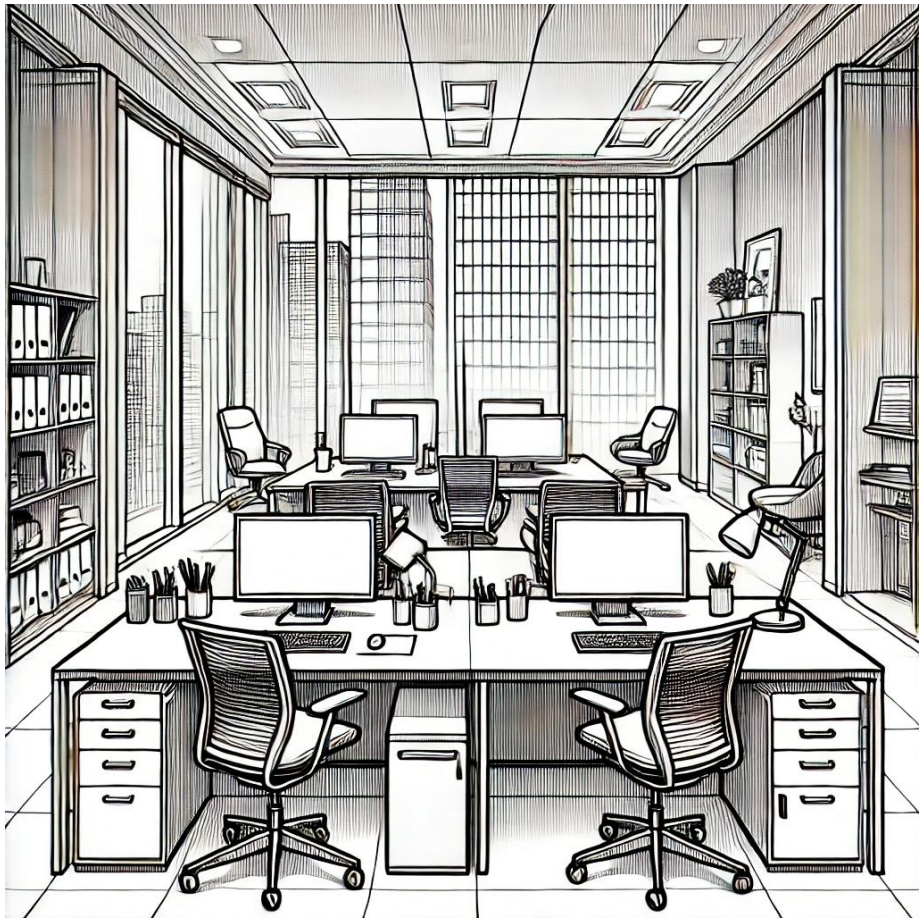
Room Occupancy Prediction

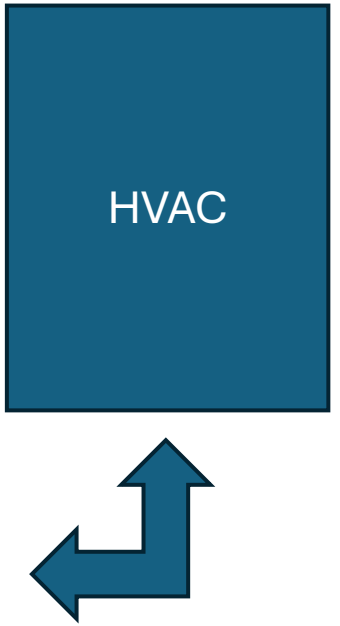
Case Study

Tim Becker – 31.3.2025

Problem Statement

- Predict office occupancy to control heating, ventilation and air conditioning (HVAC)
- Motivation is to save costs by turning off HVAC if nobody is in the room
- Minute frequency sensor data for **temperature**, **humidity**, **CO2 concentration** and **light** is given
- Data is available for approx. two weeks in February
- **Question**: Can we find patterns in the available data to predict office occupancy and thereby save costs?





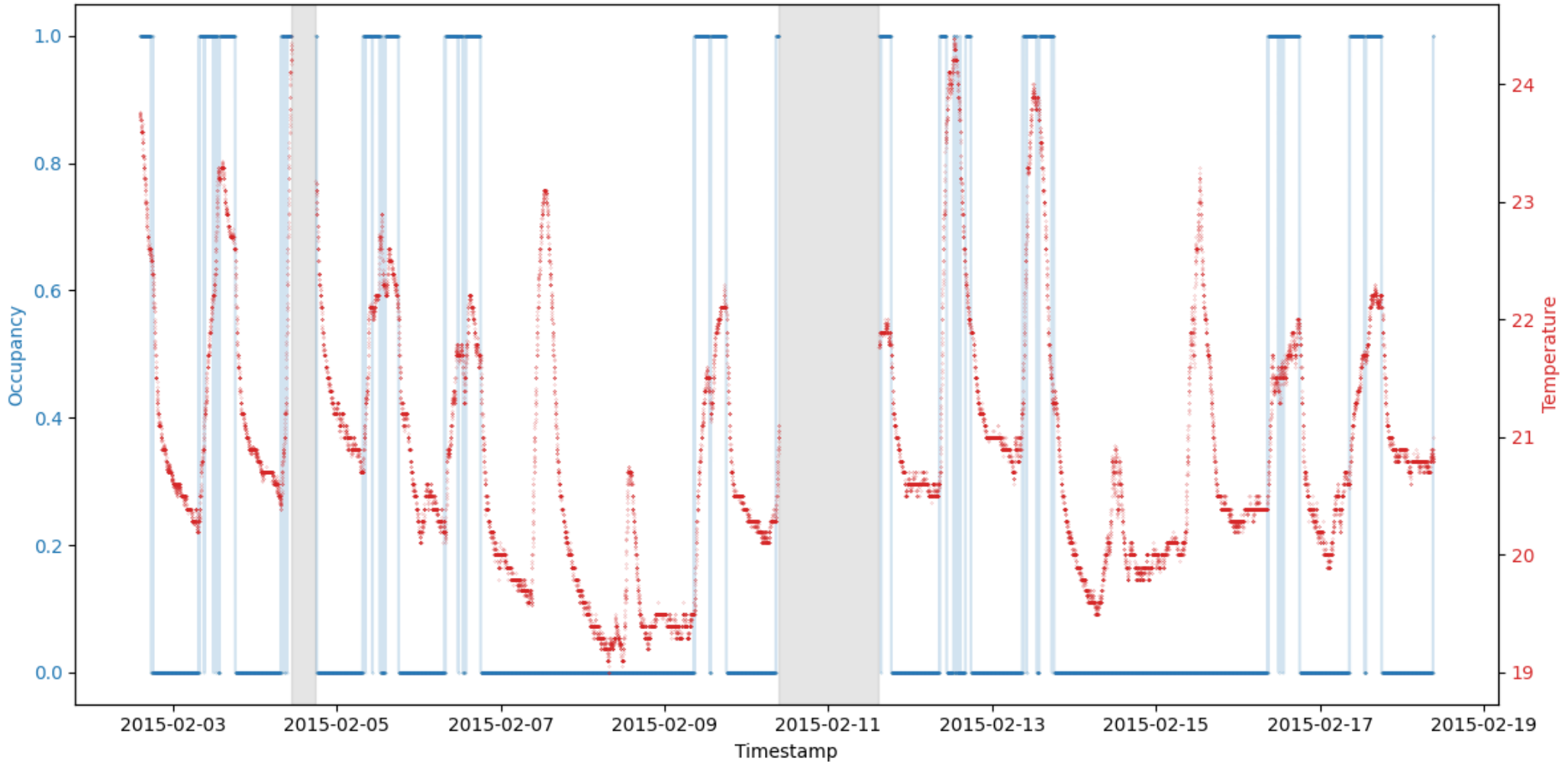
Assumptions & Comments

- We do not have any information concerning the HVAC system
  - Is it currently running and how?
  - In a real scenario, I would call someone to understand the HVAC system
- We do not know how the labels were created
- Data is only available for one shared office and for a short period
- Windows do not open or are not used in the two weeks
- Several people in the room
- No missing data

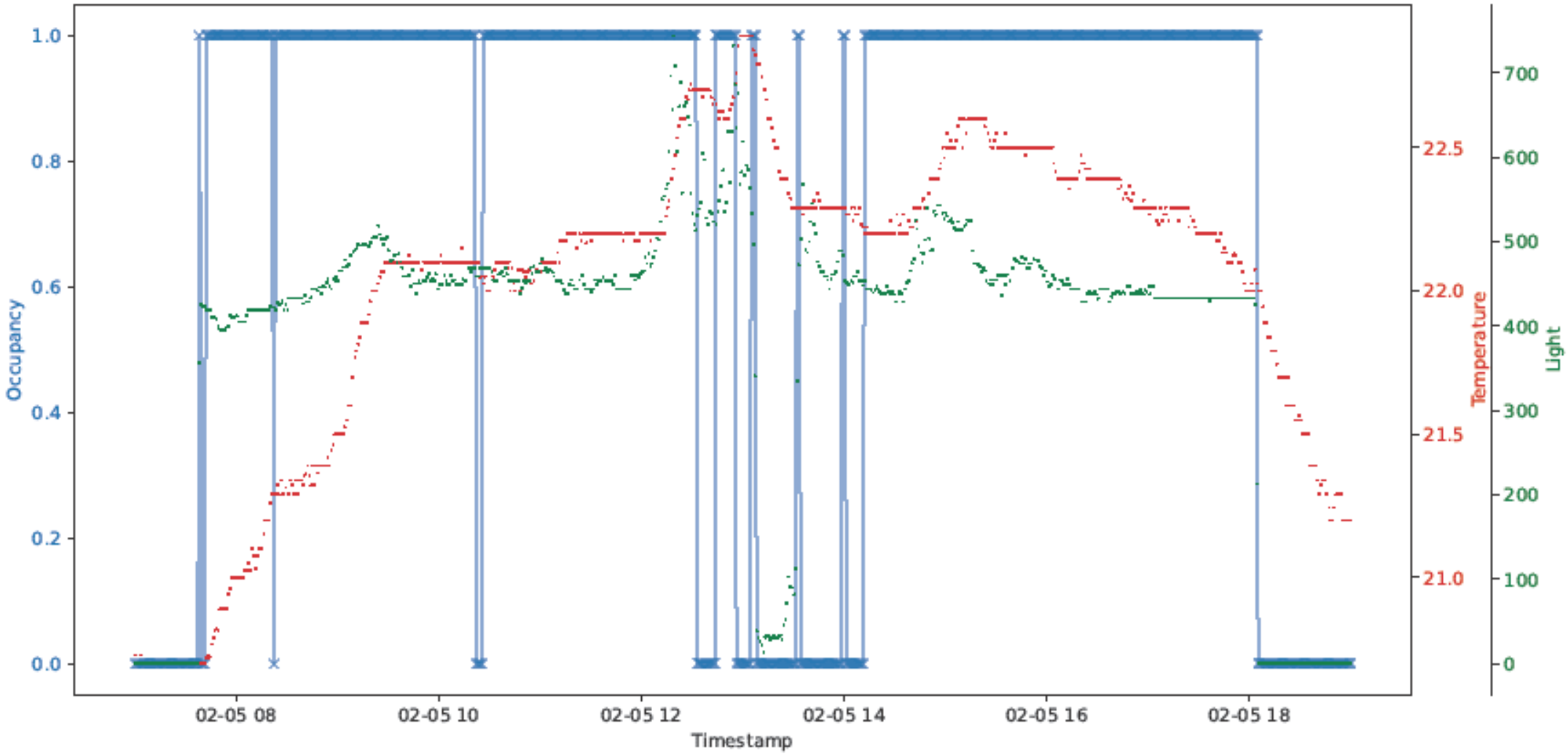
Problem-Solving Approach

1. Analyze and understand the data and the problem
2. Find a baseline and understand the potential
3. Build new features
4. Create a model to predict office occupancy
5. Evaluate the results, learn and conclude
6. Recommend way forward based on the results

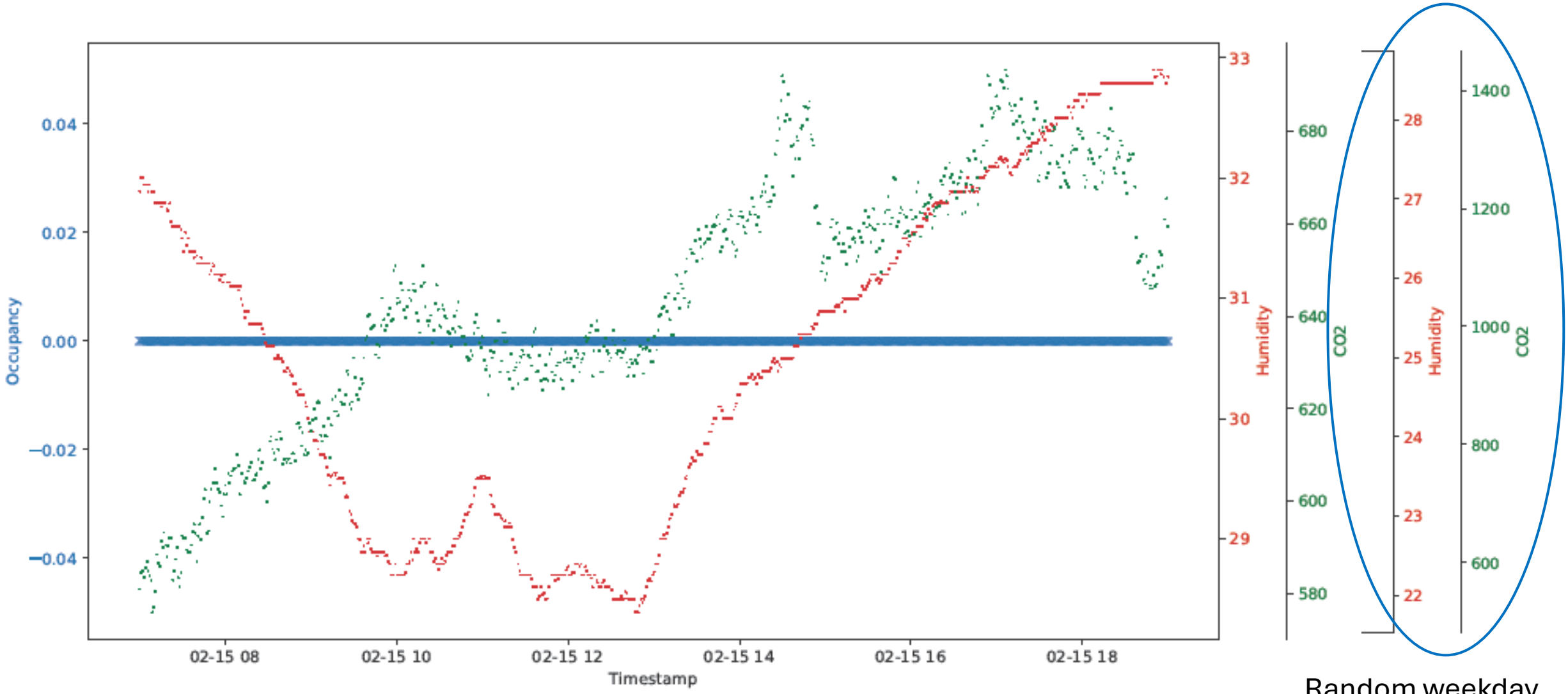
Data Analysis – Complete Period



Data Analysis – One Office Day

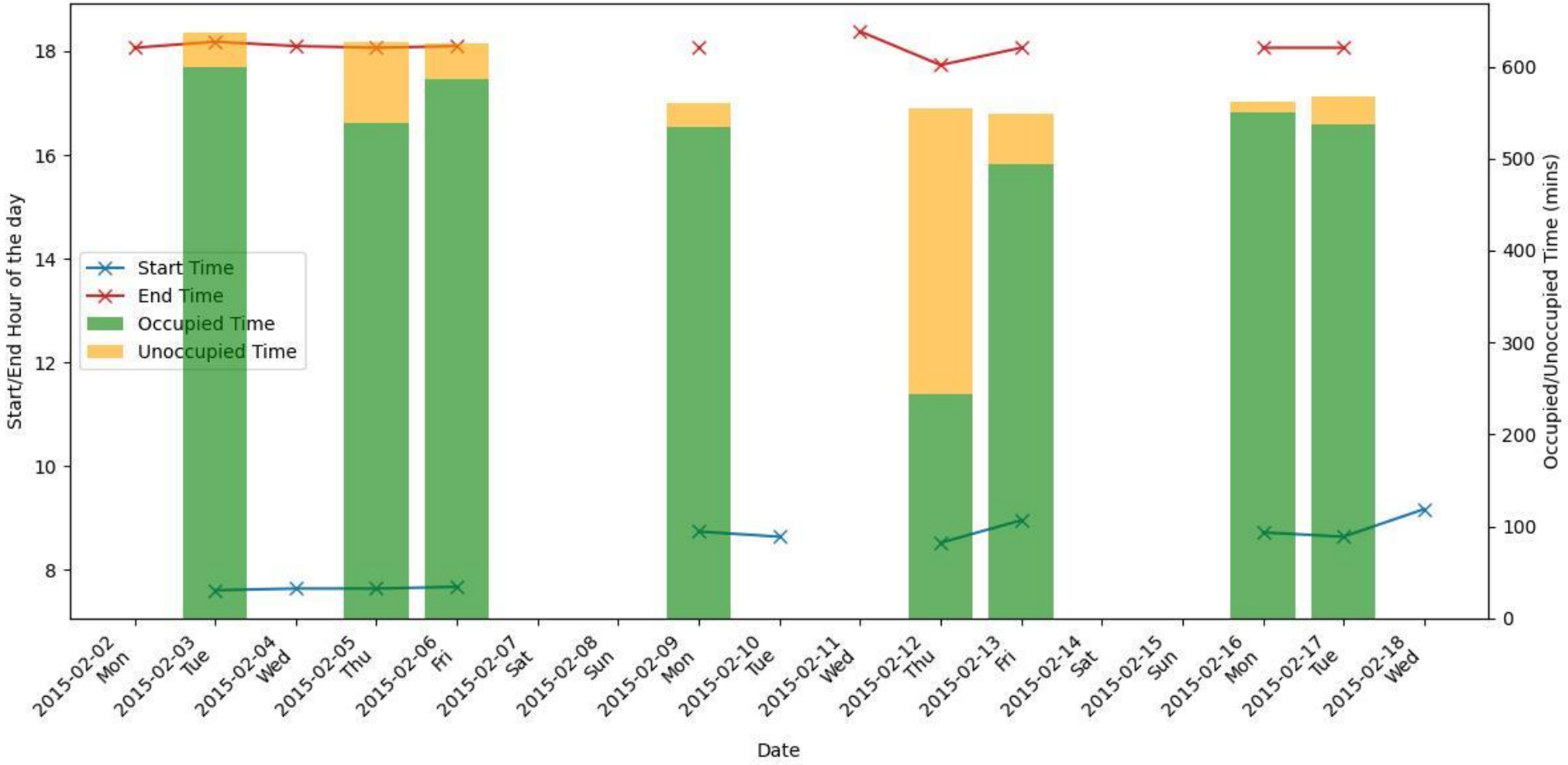


Data Analysis – Weekend

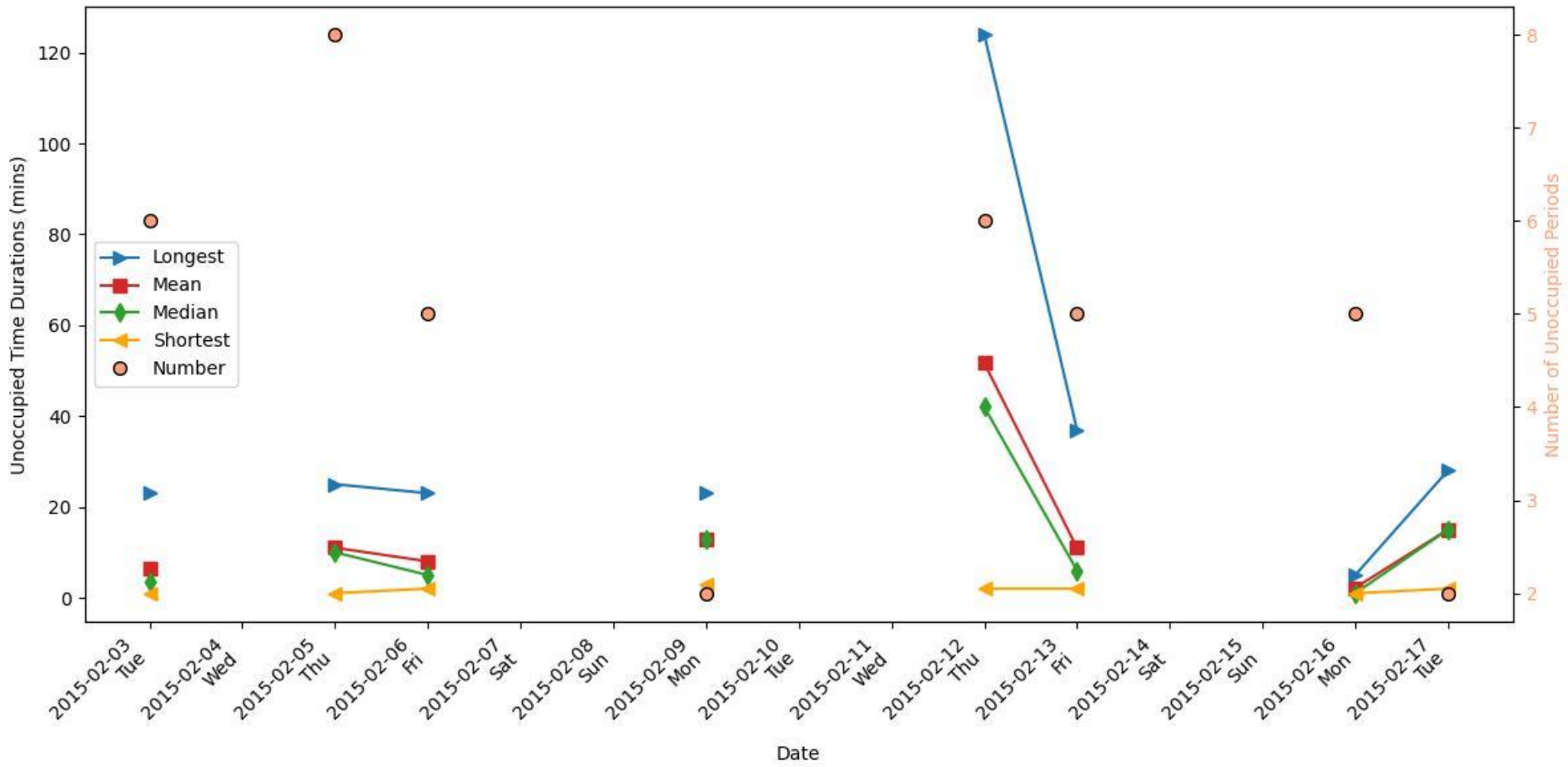


Random weekday

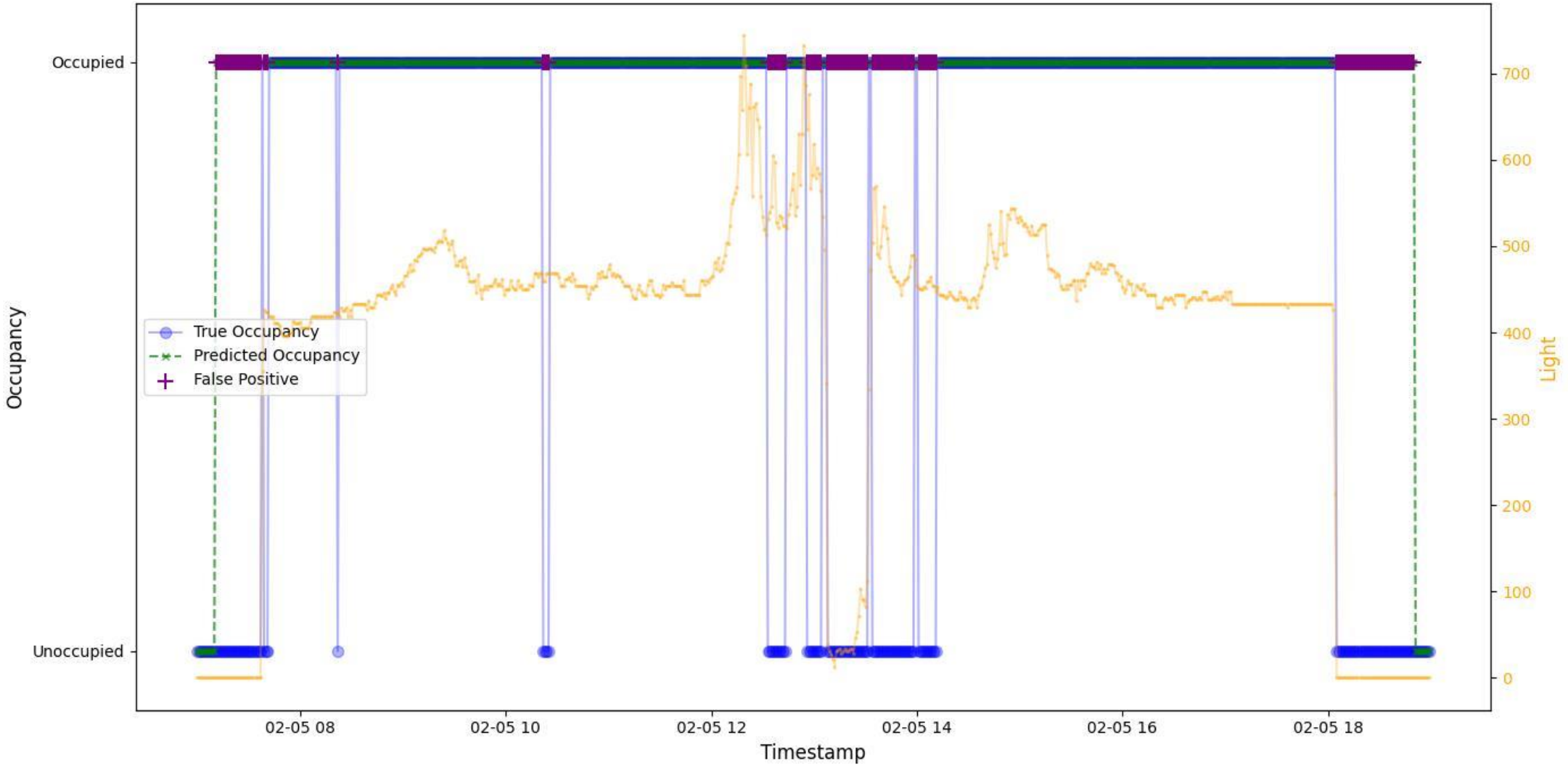
Baseline – Daily Occupancy



Baseline – Unoccupied Statistics



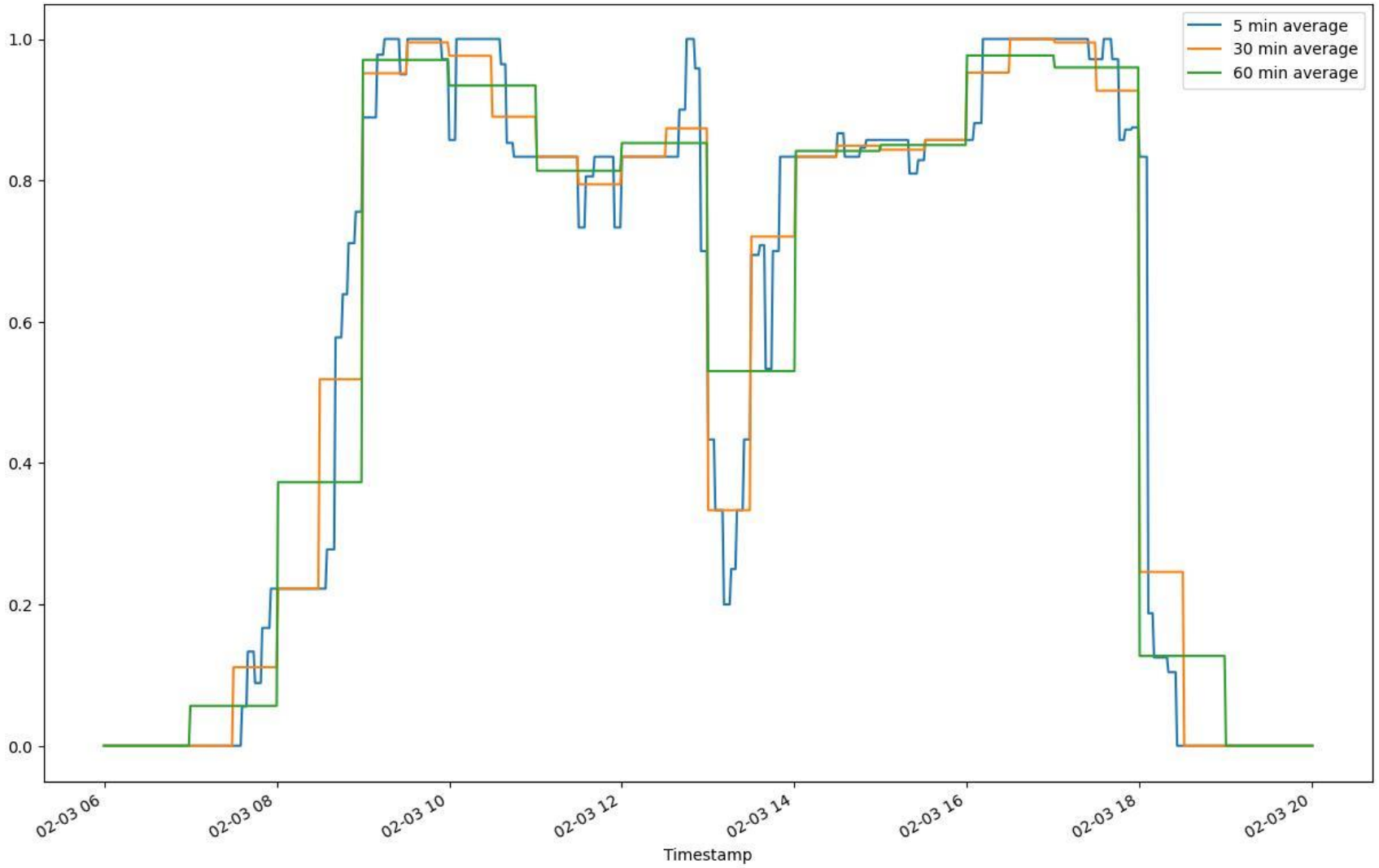
Baseline – Model



Feature Engineering – Linear Correlation

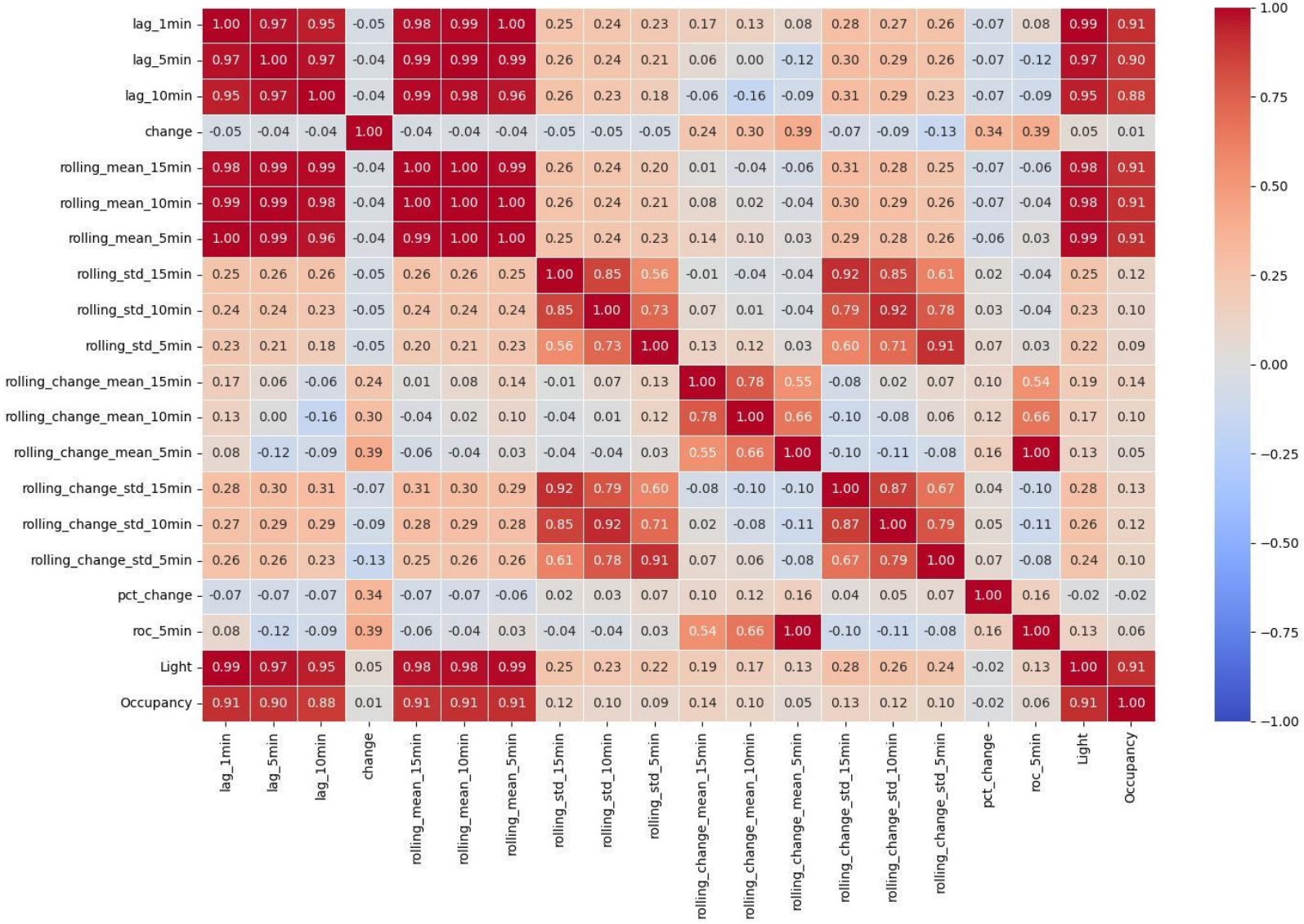
- Linear correlation
- For the given features
- Light, Temperature and CO2 have the highest values

Feature Engineering – Historic Pattern



Feature Engineering – Historic Features

- Try other features
- Light value in the past
- Average value in the past
- Change, pct change
- Mean values
- Standard devidation

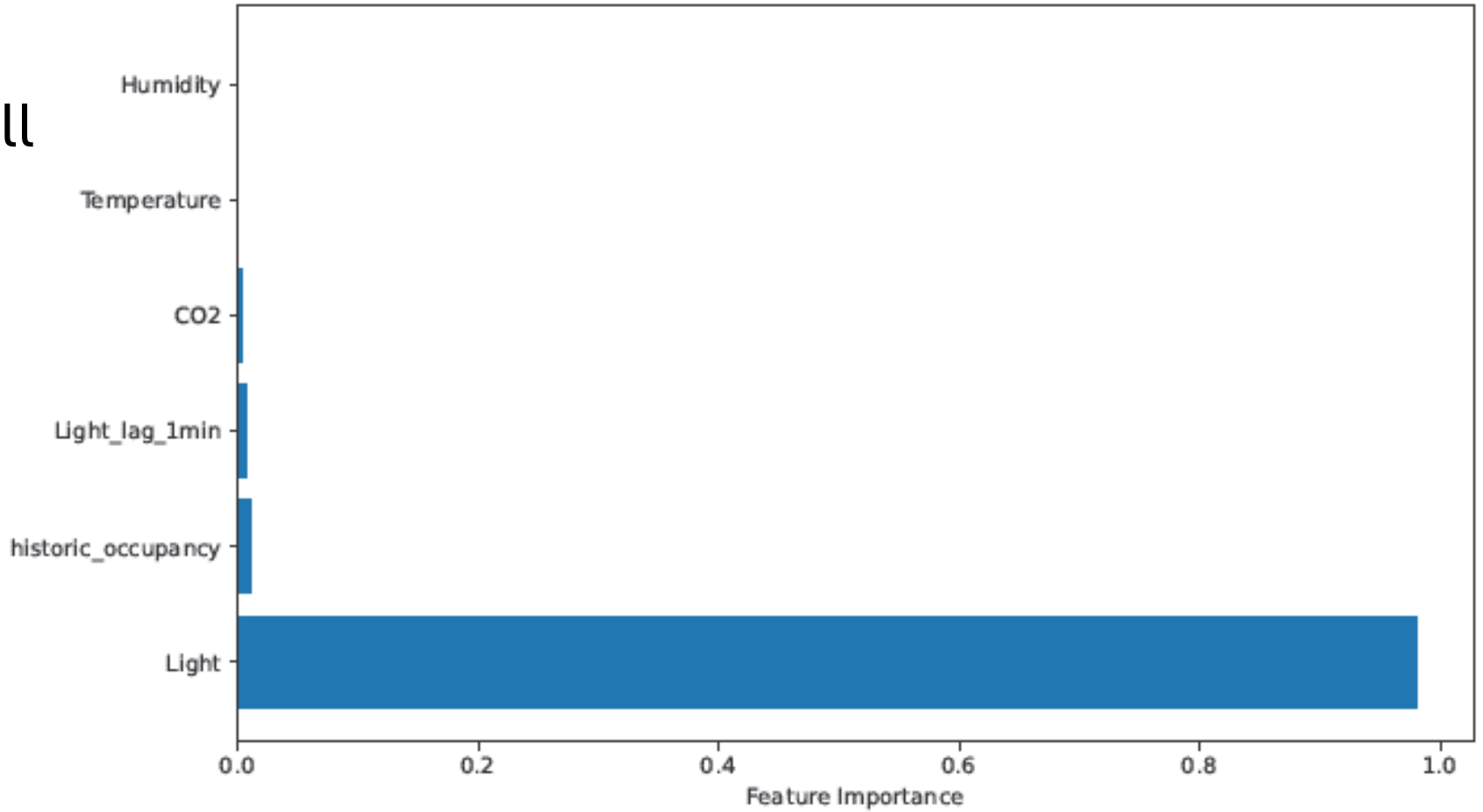


Experimental Setup

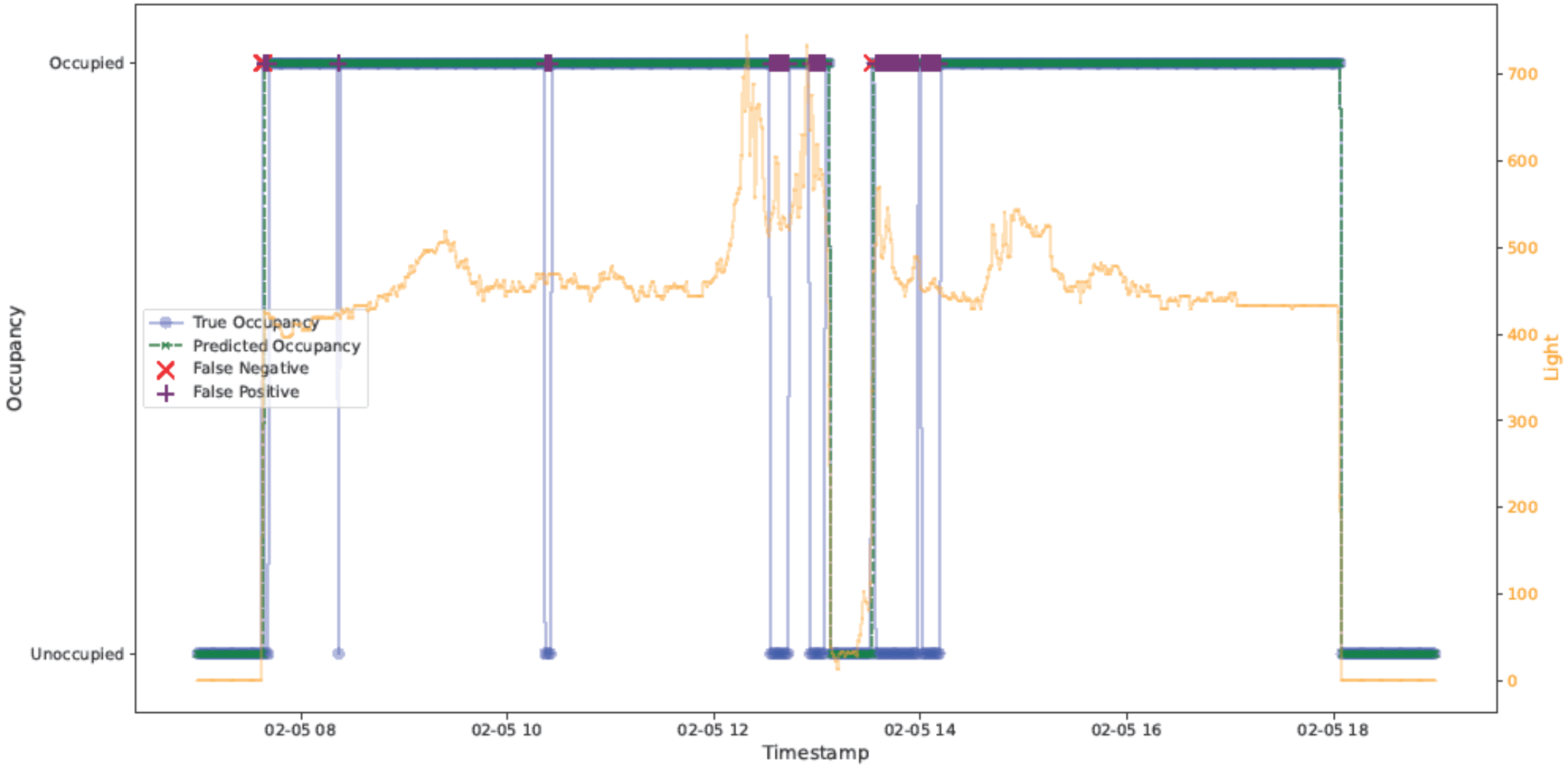
- Parameter search to find the best model and hyperparameters
  - Including new features
  - Linear model, decision tree, random forest, boosting tree
- Scale features and log transform Light and CO2
- Use cross validation with 5 splits to fit parameters
- Keep 2 complete weekdays, one weekend day and one partial day as validation dataset
- Use metrics for evaluation

Evaluation – Feature Importance

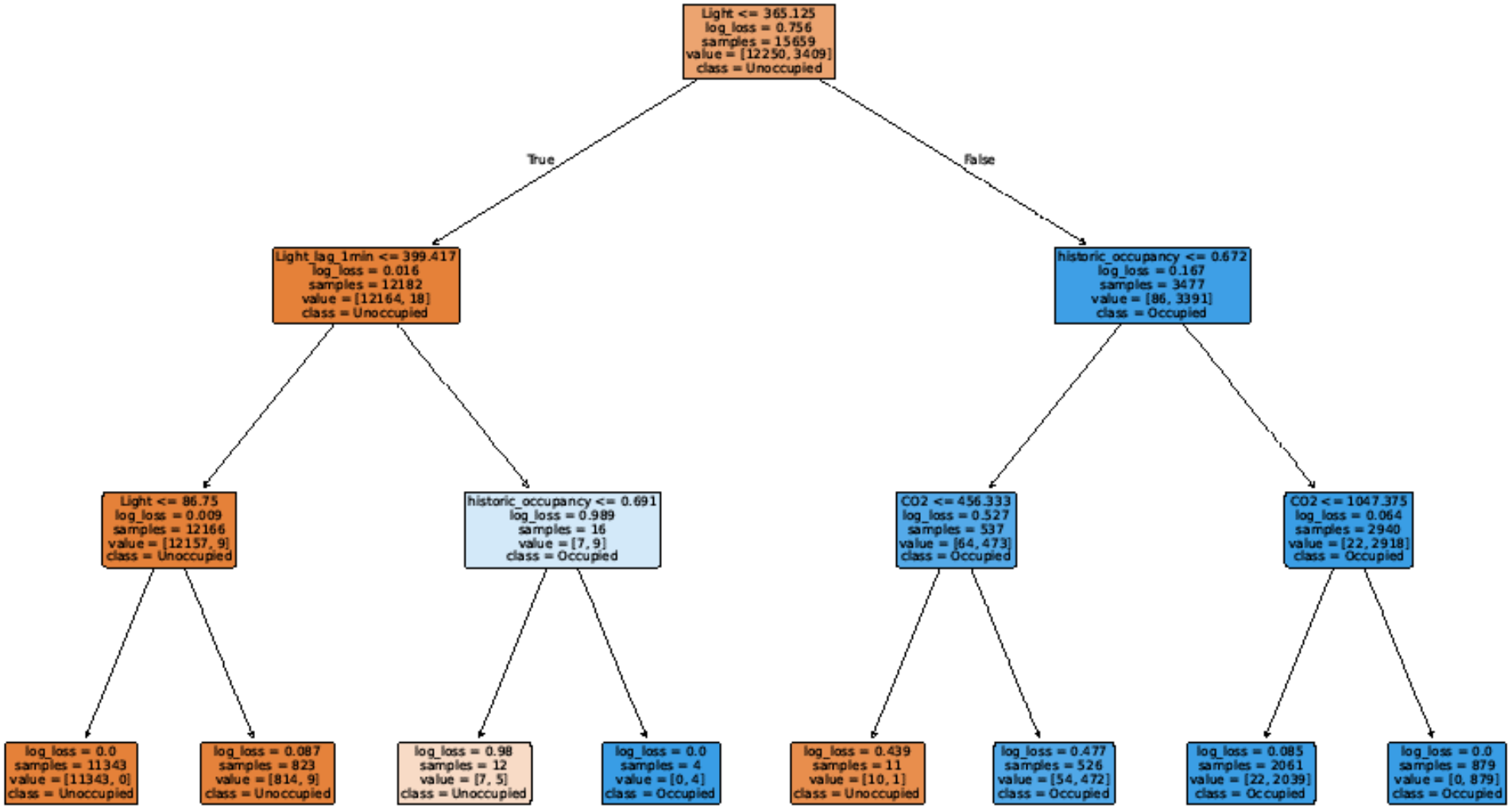
- Decision tree, but looks similar for all models
- Light is the dominant feature



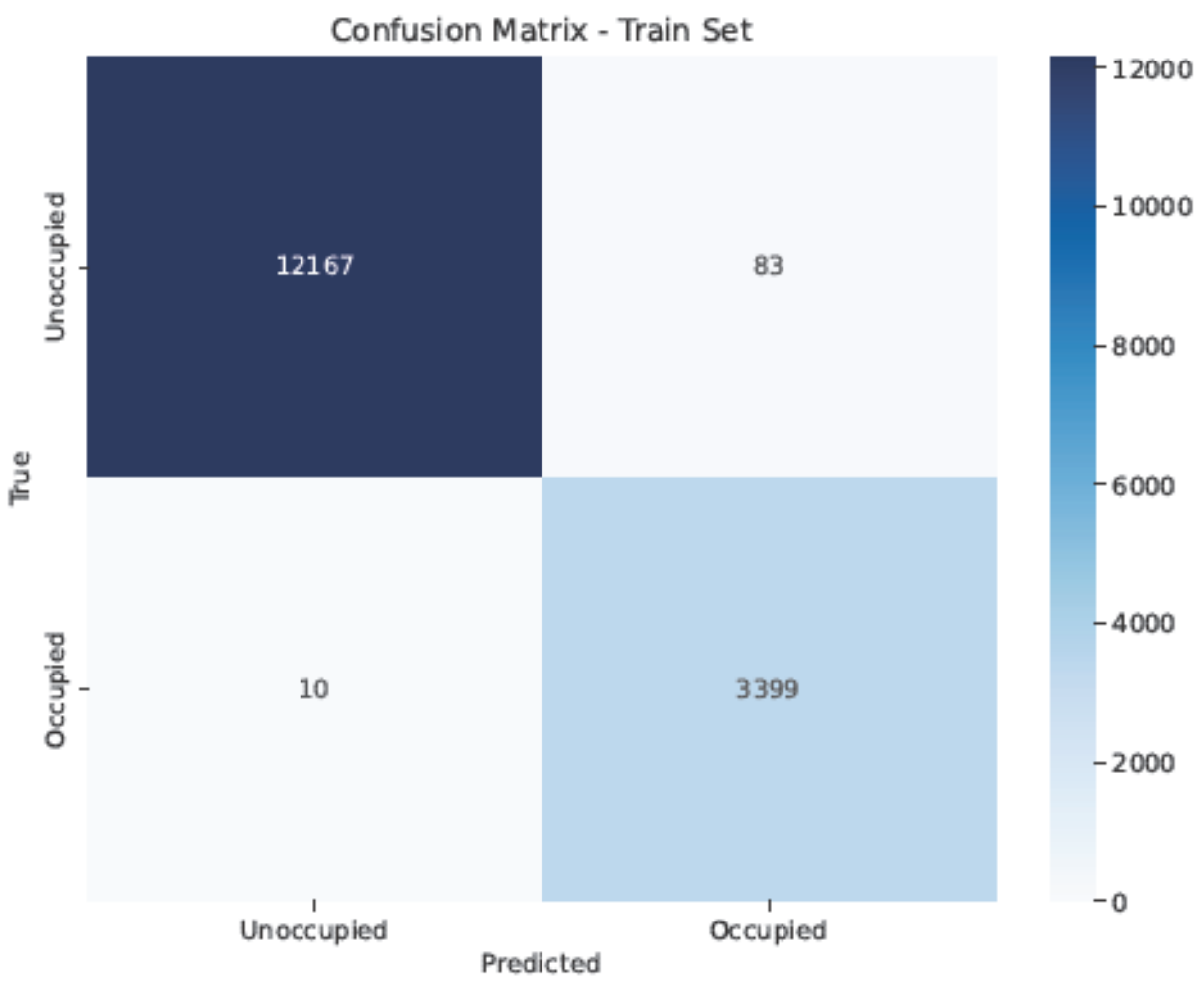
Evaluation - Predictions

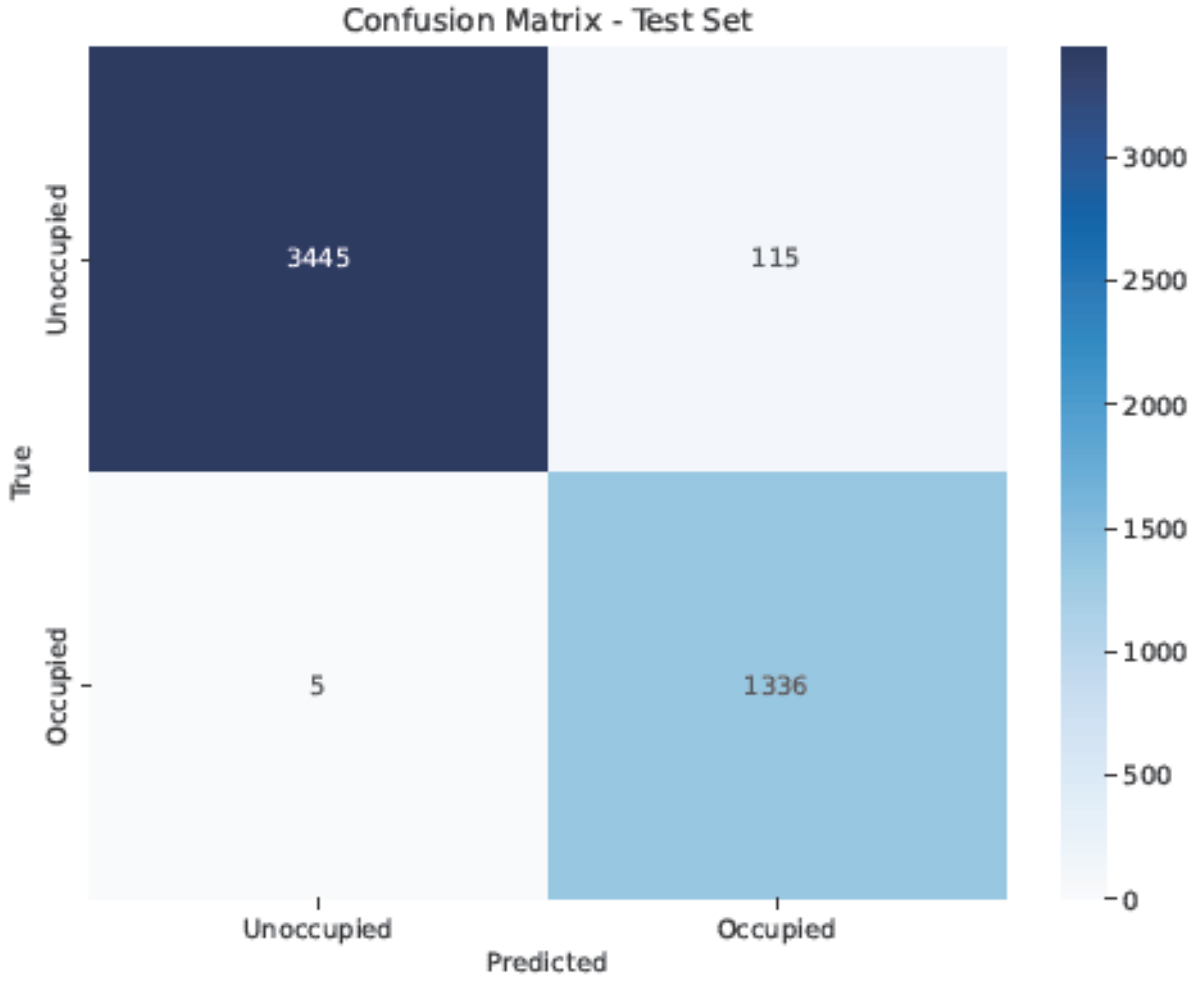


Evaluation – Tree Model



Evaluation – Scores – Decision Tree





Evaluation - Metrics

		baseline	Linear	Tree	RF	XGB
Train	accuracy	0.9	0.99	0.99	1.0	0.99
	precision	0.68	0.97	0.98	0.99	0.98
	recall	1.0	0.98	0.99	1.0	1.0
	f1	0.81	1.0	1.0	0.99	0.99
	False negative	0	23	10	13	16
	False positive	1579	117	83	46	82
Val	accuracy	0.93	0.98	0.98	0.98	0.98
	precision	0.8	0.93	0.92	0.92	0.93
	recall	1.0	0.96	0.96	1.0	1.0
	f1	0.89	0.99	0.99	0.96	0.96
	False negative	0	4	5	3	5
	False positive	331	106	115	111	107

Evaluation – Reflection

- All models perform similarly
- The model learned when the light is turned on or off
- Better than the baseline
- Unoccupied periods depend on the behavior of the employees
  - Random coffee breaks or un-regular meetings will be very difficult to predict
- Features like humidity or CO2 concentration do not contribute
  - could be because the number of people in the room changes
  - or the few days of data are insufficient for learning complex patterns
- We do have some false positives during the weekends

Possible Improvements

- We need more data
  - Data for different rooms
  - Light might not work during summer
  - Try more complex models
- Behavior might differ between days
  - Maybe people leave earlier on Friday
- Use more features
  - Coffee machine usage
  - Canteen menu
  - Outlook calendars
  - Outside weather data
  - Operating data from HVAC
- Explore time features a bit more, especially weekends, might slightly increase the score

Recommendations

- What are we trying to do?
  - Predicting the behavior of people might be challenging
  - Might cause issues with GDPR and unions
  - People will be unhappy if there are mistakes
  - Every room and its occupants are different
  - Does it make sense to stop heating, light, ventilation for short breaks?
- It might be expensive to get a lot of high-quality labeled data
- Costs of maintaining sensors and models for every room might be high
- With more data sources and longer periods, we might be able to predict longer breaks
- **Why not use something like motion sensors?**
- **Look usage/cost data of the HVAC system to find opportunities**

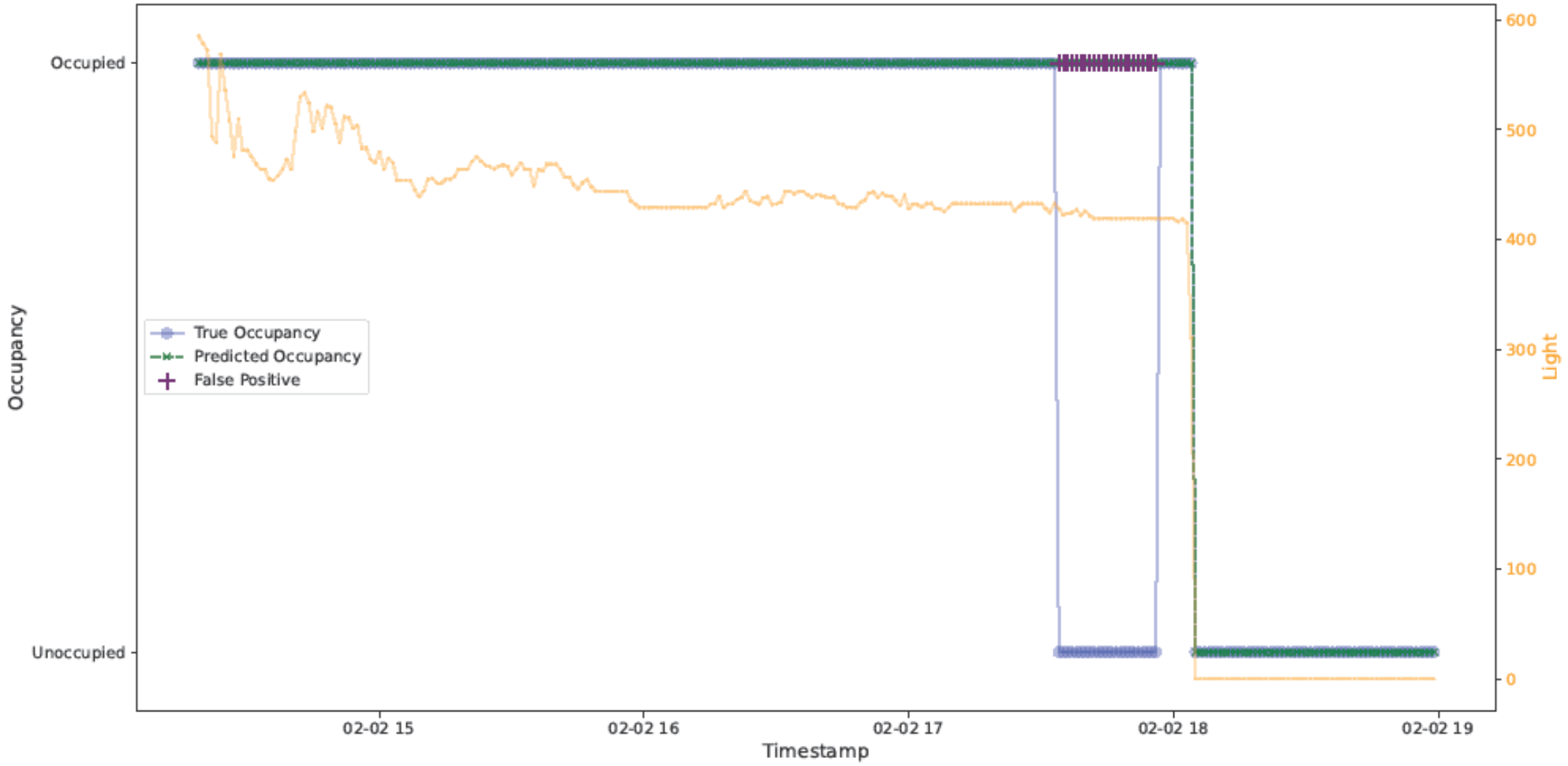
Model Deployment

- How can we integrate with HVAC system?
  - Predictions every minute with latest data
- A simple light & history model could potentially be deployed in the room together with the sensors and connected directly with HVAC
  - Advantages: speed, connection easy, employee data is not shared
  - Drawbacks: more difficult maintenance and model updates
- Deploy to the cloud
  - Advantages: easy monitoring and versioning, use existing infrastructure
  - Drawbacks: likely slower, more chances for failure, complexity

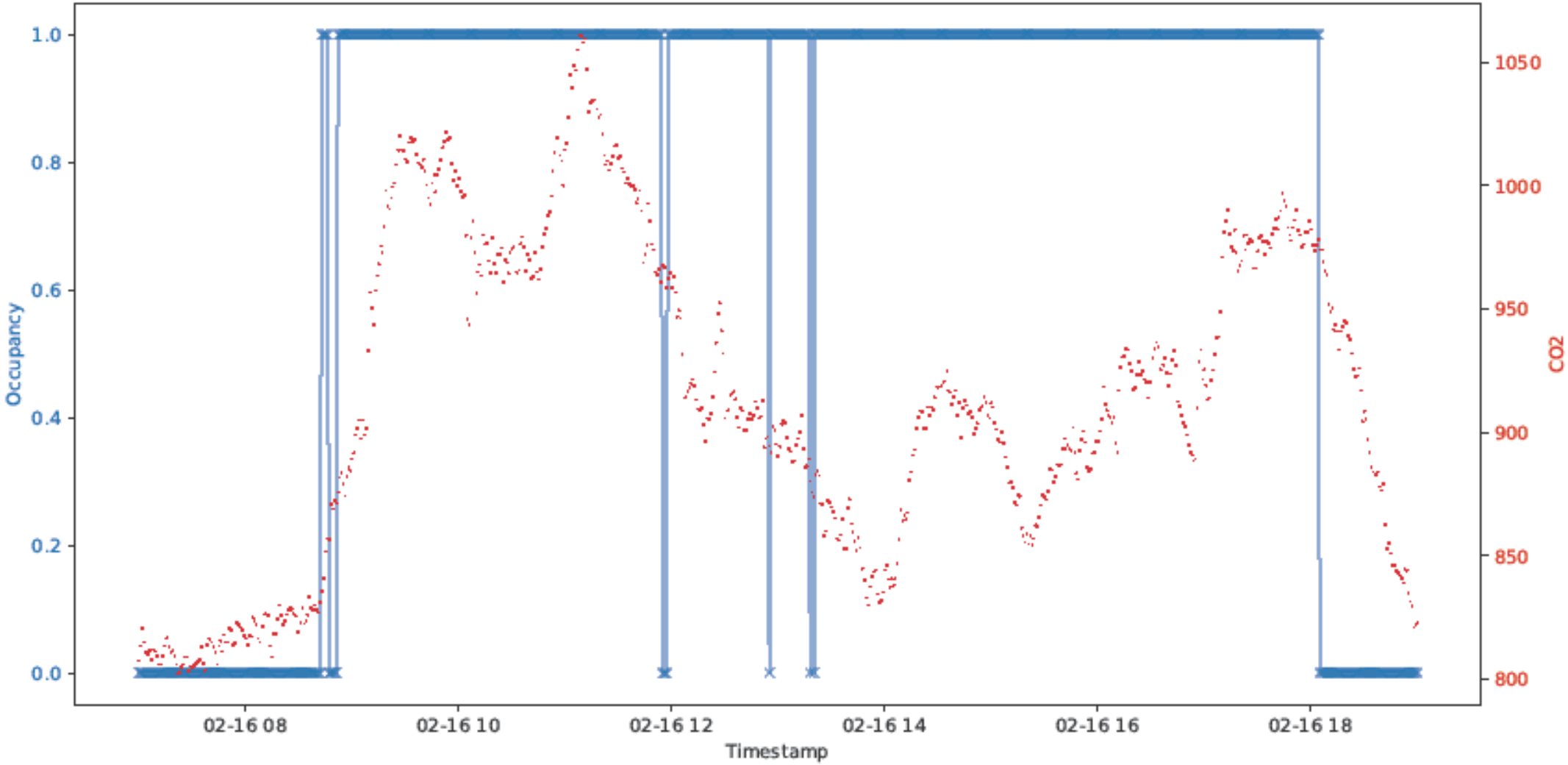
Model Monitoring

- Alarms if the infrastructure fails, e.g., heartbeat
- Track performance, but we have no labels
  - Compare input and output distributions to the past
  - Anomaly detection
  - Human feedback
  - Create labels based on daily schedule for retraining?

Appendix – Light left on



Appendix – Several people



Units

- Temperature - °C
- Relative Humidity - %
- Light - Lux
- CO2 – ppm
- HumidityRatio – kg water-vapor/kg-air
- Occupancy – 0 not ouccupied