Status Report:
UBCO MDS Capstone - Urban Data Labs

WEEK 7

Connor Lee, Claudia Nikel, Eva Nguyen, Alex Tamm

Outline

- Progress made during previous week
  - Individual logs
  - Team logs
- Current Progress
- Preliminary Results
- Database Changes
- Difficulties & Roadblocks
- Plan for next cycle

Previous Week’s Progress

Progress - Individual Work Logs

Connor

Researched model comparison methods + queried additional data + wrote code to test the various clustering methods + integrated Random Forest code into main function + got output from clustering step + integrated weather data into code

Claudia

Fixed Ridge Regression code + coded for getting average MSE for Ridge Regression part + updated github + started final report layout + reviewed clustering and Random Forest code

Alex

Finished date range analysis + code review for aggregation function + sent UDL influx-SkySpark data + started code for last step of model (pipe output to InfluxDB) + updated end-use labels to create finalized training data

Eva

Integrated Ridge Regression code into main.py + created dummy data for testing + optimized code by fixing copy slice issues and find() functions + made feature selection into a module

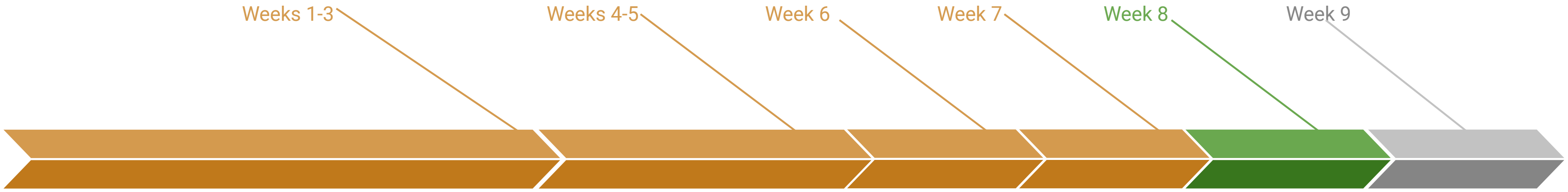
Progress - Team Work Logs

Accomplishments

- Found list of problematic data & sent it to UDL
- Expanded our Main_Pseudocode file to include expected step outputs
- Queried additional days of data
- Developed grid search code for optimizing each model
- Created more finalized training & testing data from results of electrical panel

Current Progress

Project Schedule



Weeks 1-3

Investigation & Data Prep

- Identify project objectives and key data features

- Understand data dictionaries

- Transform data for machine learning tasks

Weeks 4-5

Feature Selection/Engineering

- Research feature selection techniques

- Merge data & metadata

- Make categorical data into smaller fields

- Aggregate different values

- Identify relevant continuous & categorical features

- Create testing and training data

Week 6

Initial Modelling

- Create 3 models for each step in our project

- Run test through main.py with test dataset to get a result

Week 7

Model Tuning

- Adjust parameters of model

Week 8

Finalize Model & Visualization

- Validate & evaluate model

- Create visualization of results

- Start Final Report & Presentation

Week 9

Wrap-up

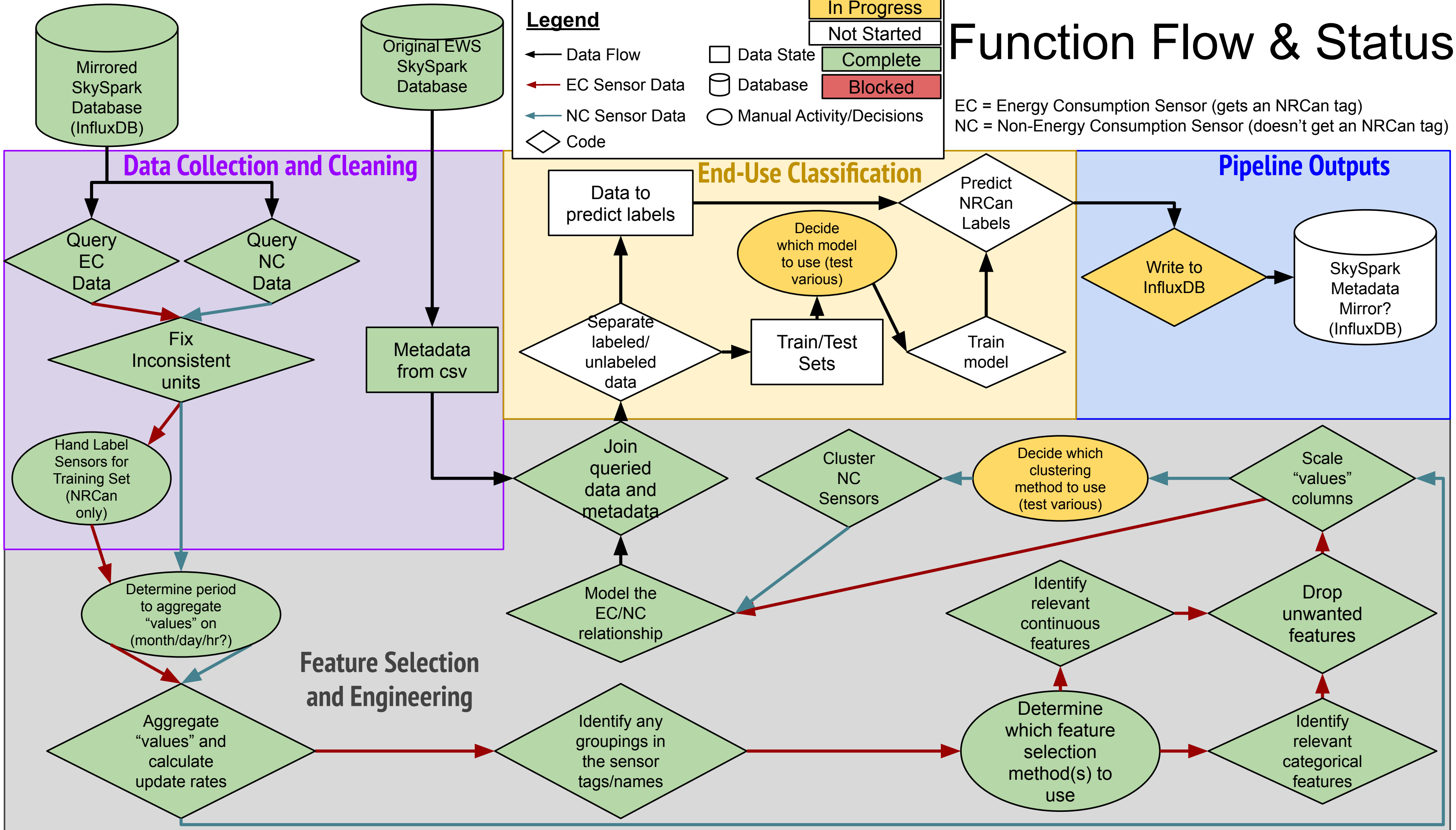
- Presentation

- Final report

- Package final code

Function Flow & Status

EC = Energy Consumption Sensor (gets an NRCan tag)
NC = Non-Energy Consumption Sensor (doesn't get an NRCan tag)



Legend

← Data Flow
← EC Sensor Data
← NC Sensor Data
◇ Code
□ Data State
Database
○ Manual Activity/Decisions

🟨 In Progress
□ Not Started
🟩 Complete
🟥 Blocked

Mirrored SkySpark Database (InfluxDB)

Original EWS SkySpark Database

Data Collection and Cleaning

Query EC Data

Query NC Data

Fix Inconsistent units

Metadata from csv

Hand Label Sensors for Training Set (NRCan only)

End-Use Classification

Data to predict labels

Decide which model to use (test various)

Predict NRCan Labels

Separate labeled/ unlabeled data

Train/Test Sets

Train model

Pipeline Outputs

Write to InfluxDB

SkySpark Metadata Mirror? (InfluxDB)

Join queried data and metadata

Cluster NC Sensors

Decide which clustering method to use (test various)

Scale "values" columns

Determine period to aggregate "values" on (month/day/hr?)

Model the EC/NC relationship

Identify relevant continuous features

Drop unwanted features

Feature Selection and Engineering

Aggregate "values" and calculate update rates

Identify any groupings in the sensor tags/names

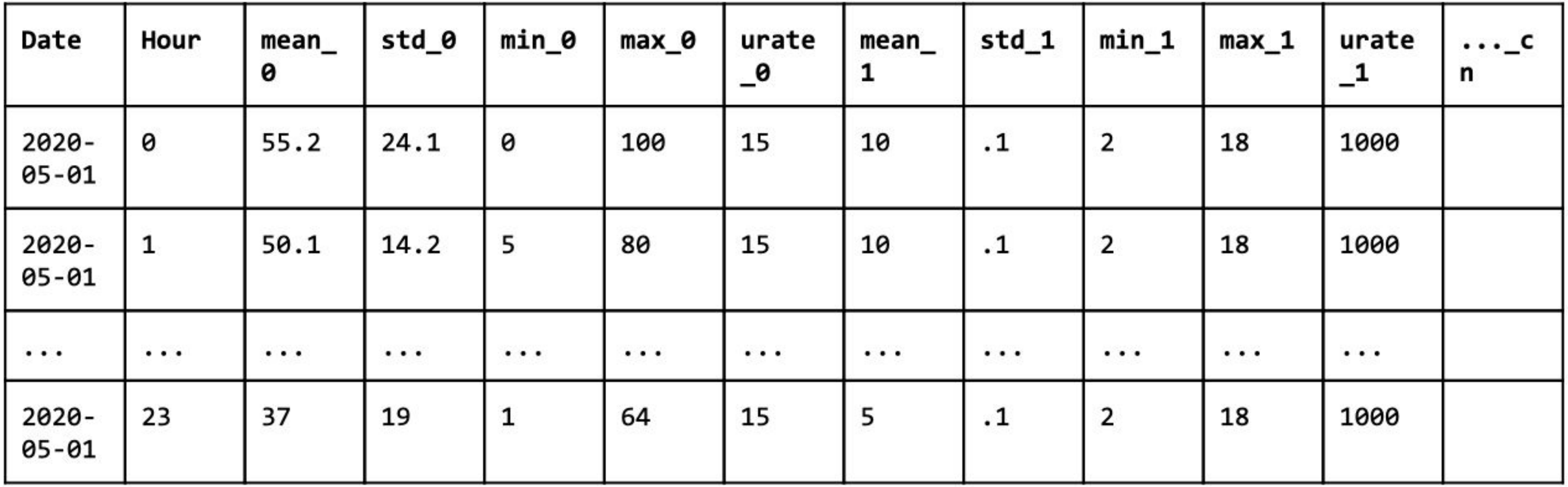
Determine which feature selection method(s) to use

Identify relevant categorical features

Preliminary Results

Clustering

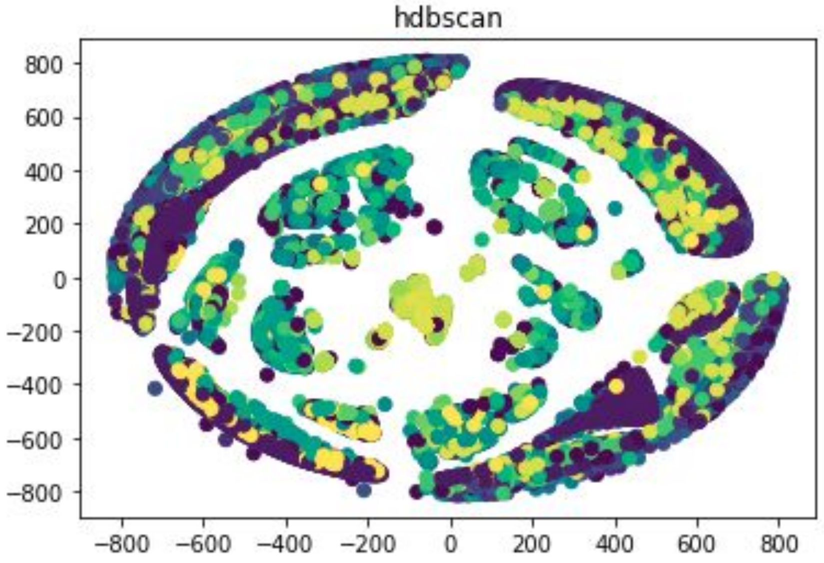
→ Cluster NC Sensors



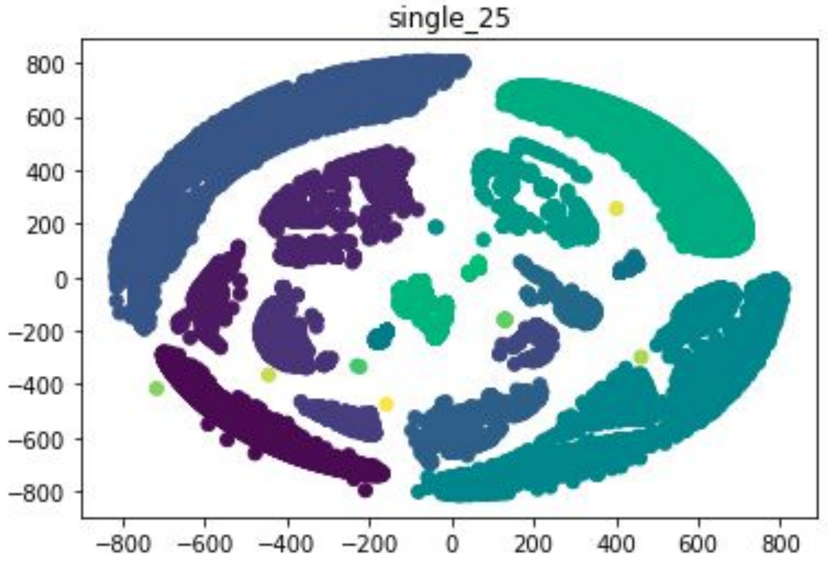
Date	Hour	mean_0	std_0	min_0	max_0	urate_0	mean_1	std_1	min_1	max_1	urate_1	..._cn
2020-05-01	0	55.2	24.1	0	100	15	10	.1	2	18	1000	
2020-05-01	1	50.1	14.2	5	80	15	10	.1	2	18	1000	
...	...	...	...	...	...	...	...	...	...	...	...	
2020-05-01	23	37	19	1	64	15	5	.1	2	18	1000	

Clustering

→ Cluster NC Sensors



✗

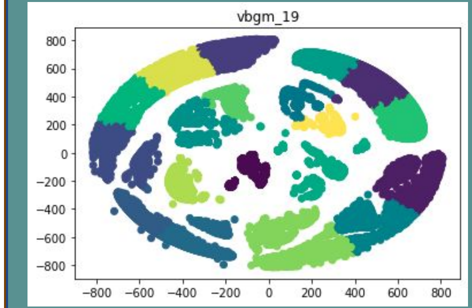


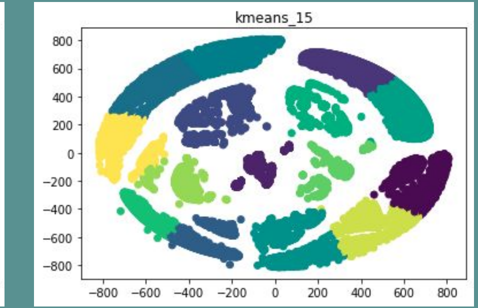
✓

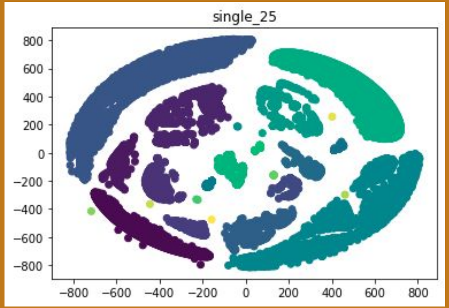
Clustering

→ Cluster NC Sensors

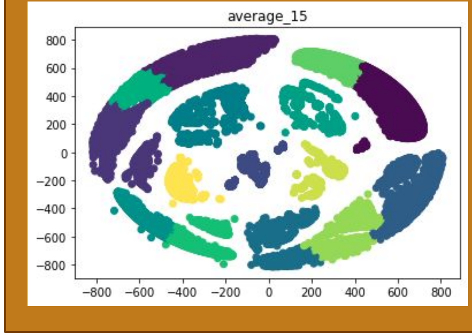
Non-Agglomerative

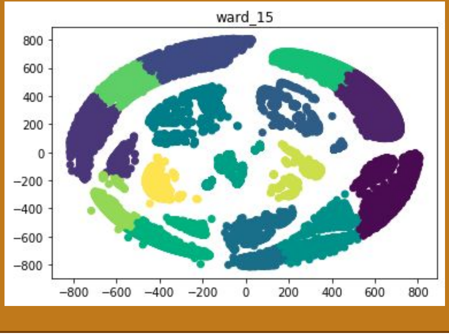

vbgm_19

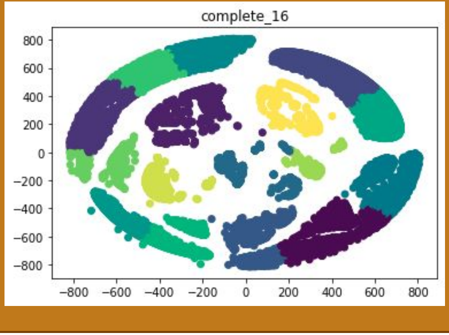

kmeans_15

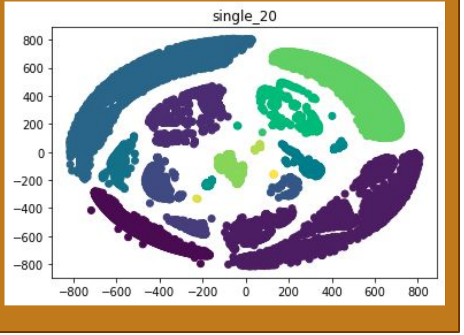

single_25

Agglomerative


average_15

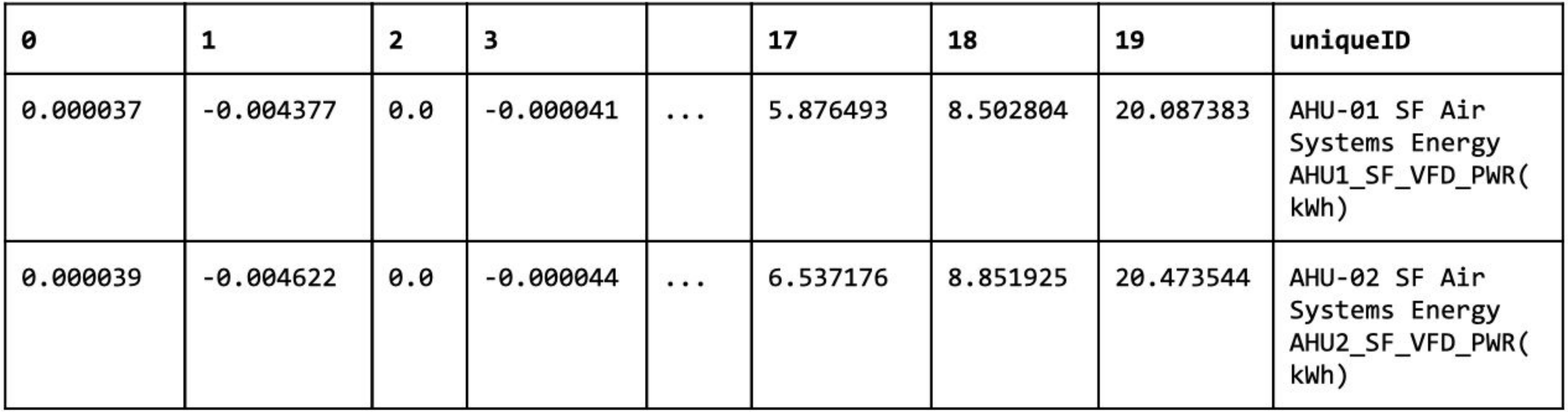

ward_15


complete_16


single_20

Regression

→ Model EC/NC Relationship



0	1	2	3		17	18	19	uniqueID
0.000037	-0.004377	0.0	-0.000041	...	5.876493	8.502804	20.087383	AHU-01 SF Air Systems Energy AHU1_SF_VFD_PWR(kWh)
0.000039	-0.004622	0.0	-0.000044	...	6.537176	8.851925	20.473544	AHU-02 SF Air Systems Energy AHU2_SF_VFD_PWR(kWh)

Coefficients from Ridge Regression for each sensor

Supervised Model

→ Predicting End-Use Labels

Confusion Matrix and Performance Metrics For The Full Date Range

End-Use Labels on Test Set	00	01	02	03	04	05
00_HEATING_SPACE_AND_WATER	[[10	0	0	1	0	0]
01_SPACE_COOLING	[0	3	0	0	0	0]
02_HEATING_COOLING_COMBINED	[0	0	7	0	0	0]
03_LIGHTING_NORMAL	[2	0	0	7	0	0]
04_LIGHTING_EMERGENCY	[0	0	0	0	2	0]
05_OTHER	[0	0	0	0	0	4]]

```
accuracy: 0.9166666666666666
precision: 0.9178240740740742
recall: 0.9166666666666666
f1: 0.9160272804774083
logloss: 0.29794974927791823
```

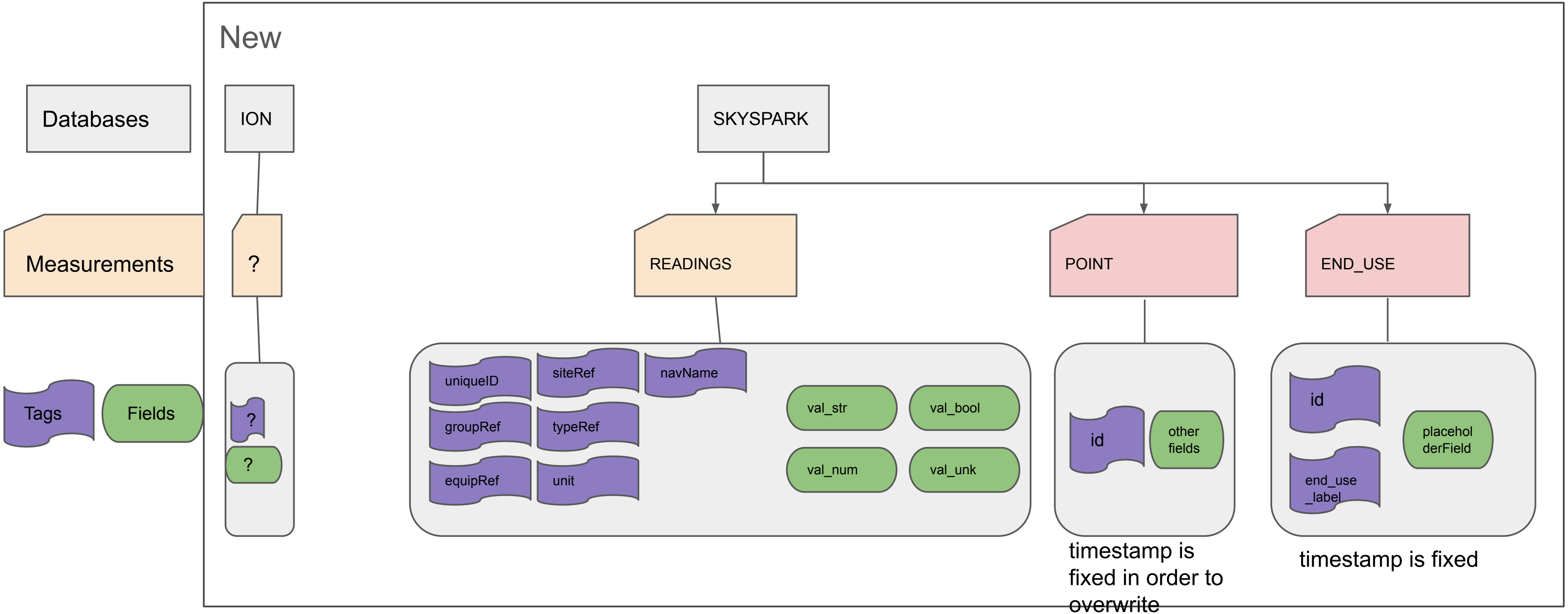
Supervised Model

→ Predicting End-Use Labels

Confusion Matrix and Performance Metrics For The Full Date Range

Database Changes

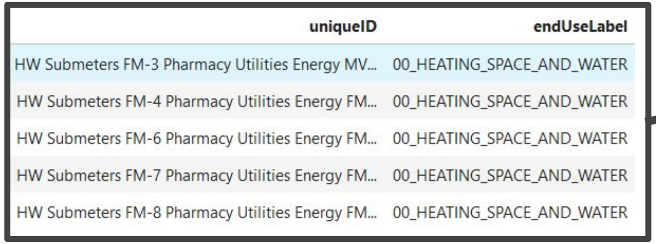
SkySpark v7 (InfluxDB)



timestamp is fixed in order to overwrite

timestamp is fixed

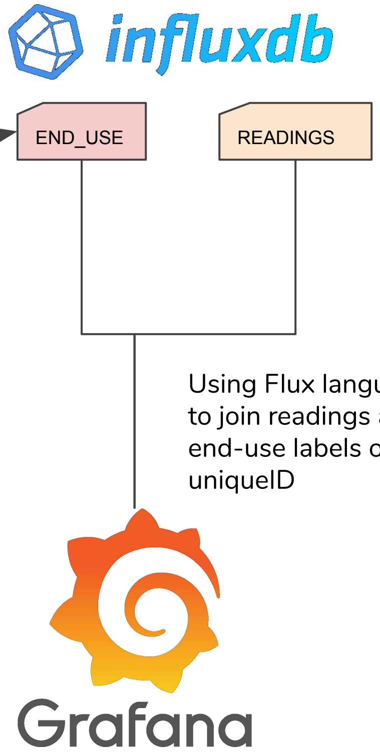
Visualization Pipeline



uniqueID	endUseLabel
HW Submeters FM-3 Pharmacy Utilities Energy MV...	00_HEATING_SPACE_AND_WATER
HW Submeters FM-4 Pharmacy Utilities Energy FM...	00_HEATING_SPACE_AND_WATER
HW Submeters FM-6 Pharmacy Utilities Energy FM...	00_HEATING_SPACE_AND_WATER
HW Submeters FM-7 Pharmacy Utilities Energy FM...	00_HEATING_SPACE_AND_WATER
HW Submeters FM-8 Pharmacy Utilities Energy FM...	00_HEATING_SPACE_AND_WATER

Output Dataframe From
Classification Model Prediction

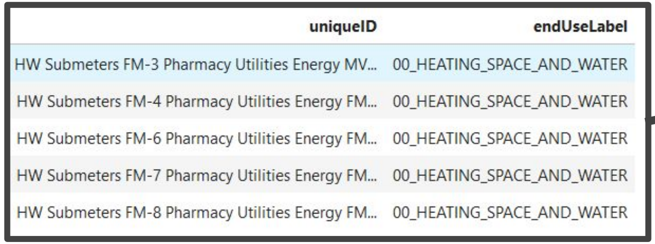
Write/Update Using
influxdb-python
package



Using Flux language
to join readings and
end-use labels on
uniqueID

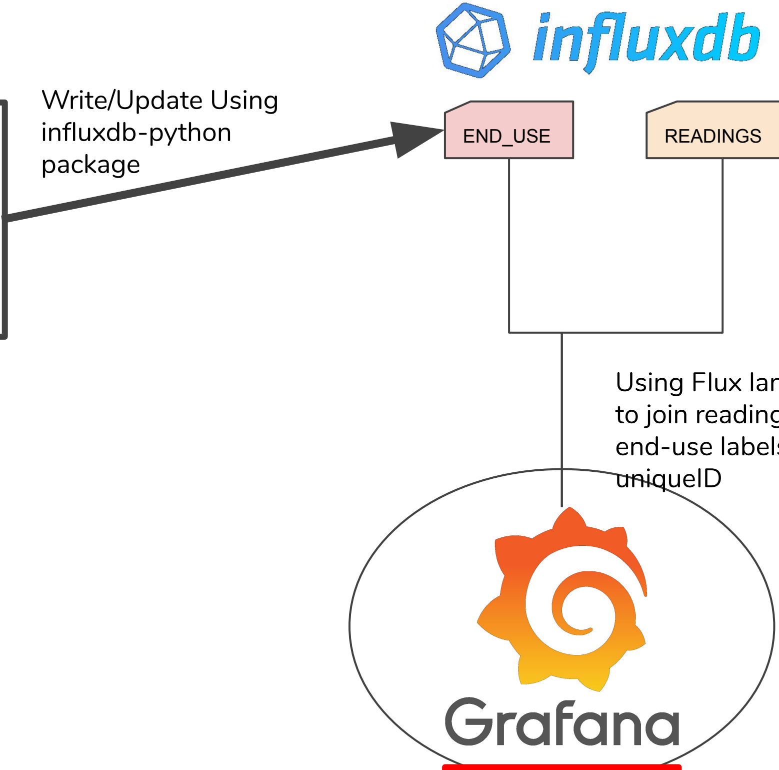
Difficulties & Roadblocks

Visualization Pipeline - Difficulties



uniqueID	endUseLabel
HW Submeters FM-3 Pharmacy Utilities Energy MV...	00_HEATING_SPACE_AND_WATER
HW Submeters FM-4 Pharmacy Utilities Energy FM...	00_HEATING_SPACE_AND_WATER
HW Submeters FM-6 Pharmacy Utilities Energy FM...	00_HEATING_SPACE_AND_WATER
HW Submeters FM-7 Pharmacy Utilities Energy FM...	00_HEATING_SPACE_AND_WATER
HW Submeters FM-8 Pharmacy Utilities Energy FM...	00_HEATING_SPACE_AND_WATER

Output Dataframe From Classification Model Prediction



Write/Update Using influxdb-python package

Using Flux language to join readings and end-use labels on uniqueID

Any side-effects of running with influxDB1.7.8?

Where to host?

Difficulties

- Code collaboration
- Choosing a time efficient test method
- Writing efficient code
- Modifying code to include try and except statements
- Making code for each step in the model cohesive
- Joining data on a tag in influxDB

Tasks for Next Cycle

Tasks for the Next Weekly Cycle

1. Finish tuning model
2. Start writing final report
3. Start creating final presentation
4. Create Jupyter notebook on how to use the main.py & modules
5. Create placeholder dashboard (visualization)

Questions