Weekly Sprint Planning

2020-05-08 /10:00-11:30 / Zoom

THEMES	WEEKS	DATES	GOALS
Investigation and Data Prep	3	27 April - 14 May	Identify project objectives and key data features + understand data dictionaries + transform data for machine learning tasks.
Model	1	15 - 21 May	Develop a classification model to apply group tags to end-uses for the Pharmacy building.
Model	1	22 - 28 May	Validate and evaluate models.
Scale + Analysis	2	29 May - 11 June	Expand the model to other UBC buildings (if time permits) + complete user-acceptance testing of model + identify conclusions + create visualizations of results + complete user-acceptance testing of dashboards + UBC mid-term presentation
Wrap-Up	2	12 - 26 June	Final report + package final code + UDL final presentation + UBCO final presentation
Total Weeks	**9**		

1. What was our goal/theme from last week?
   Goal: Investigation and Data Prep

2. Which tasks did we complete?
   - **Draft Project Proposal**
   - **Prepping Tuesday's Presentation**
   - **Export Data to Master Data Set**
   - **Develop Project Overview Flowchart**
   - **Develop Workflow Document**
   - **Finalize GitHub Workflow Plan**
   - **Initial data exploration**
   - **Initial research of project solutions**

3. Was there anything stopping us from finishing specific tasks?
   - No

4. What tasks are still in progress?

- Research Problem Solutions
- Data Exploration
- Defining the Project Problem
- Explore Feature Extraction
- Finalize Proposal

5. Are there any changes that need to be made?
    - Make Jira tasks more specific
    - Assigning tasks to update in Jira
6. What is our goal/theme for this week?

Goal: Selecting features, transforming data for ML, and pipeline implementation

7. What tasks need to be added/replenished to the Backlog?
    - Prep presentation for Tuesday
    - Refer to Potential Tasks section below

8. What tasks are most important and should be pulled from Backlog to In progress?
        - Select Relevant Features
        - Set up data flow connections
        - Transform Data for ML task
        - Prep presentation for Tuesday

9. Are there any dependencies between In Progress tasks?
    a. If so, how will that be organized?
    - Confirming the project problem (UDL-???) and Set Up Data Flow Connections (UDL-23)
        - Talk to Jiachen in today's meeting to get some clarity on the specific definition
            - If he says that this is 100% what we should be doing then we can move forward
        - Start with simple method (querying) and scale up as needed

10. Who is going to be assigned to which tasks and update in Jira?

Person	In progress Tasks	New Tasks
Claudia	●	●
Connor	●	●
Eva	●	●
Alex	●	●

Potential Tasks:

- **Prep for Tuesday's Presentation**
    - Create template (Claudia)
    - 1 page summary of group progress (Claudia)
- **Confirming the project problem/project objectives**
    - Add to Aims & Objectives requirement of creating another influxdb that has a table with unique id and NRCan classification columns. This new influxdb table will join by unique id to SkySpark db to feed into a Grafana dashboard (All, at meeting with UDL))
        - Not required for Friday's proposal submission
    - Confirm with Mike and Jiachen (All)
- **Research AWS/TBD connection to stream data from SkySpark db and write to another influxdb**
    - Research what other tools can be used (All)
    - Test connection from SkySpark influxdb (Later)
    - Test writing to another influxdb (Later)
    - Set up the two-way connections (Later)
- **Develop a tool to identify missing information**

    - Outline what is meant by missing information (when the server is down, when one of the 5 identifying fields are null, etc.) (Alex/Connor)
    - Develop the code to identify when the server is down and count is 0, or when the server is down and unable to query (Alex/Connor)
    - Develop the code to identify when any fields are null or any rows within a field are null (was “omit” an identifier?) (Alex/Connor)
- **Develop a tool to populate missing information**
    - Research our approach to populating - if we want to aggregate over several readings by unique ID, or another alternative learned in data wrangling course (Eva/Claudia)
    - Develop the code to populate missing data when server is down (Eva/Claudia)
    - Develop the code to populate missing data when any fields are null or any rows within a field are null (Eva/Claudia)
- **Transform data for machine learning task**
    - Review Arthur’s comments regarding what new building data will look like and research how to prepare for that (waiting on Jiachen’s meeting with Arthur) (Eva/Claudia)
    - Scale units to be within the same range and decide if we want to convert measurements to the same scale (Pa and kPa) (Connor)
    - Encode and index non-numerical fields (Connor)
    - Vectorize the data (Connor)
    - Research different distance measures for mixed data (continuous and categorical) and how to implement them (mostly how to implement at this point, Gower’s distance seems to be the only viable option from what I have read up to now) (Connor)
    - Aggregate the data (do we want every timestamp, or aggregate by a different interval - hour, hour half, etc.) (Alex)
- **Identify NRCan's classifications**
    - F/U with Mike and Jiachen if the list Connor found suffices and/or would they like more granular classifications (Eva)

    - If more granular, work with Mike and Jiachen to get those classifications, and create a master classifications table (Eva)
- **Identify relevant features**
    - Research feature selection techniques
    - Review sensor metadata
    - Decide on a feature selection technique (one paper suggested OLS then forward selection)
    - Implement feature selection technique
    - Or, use all fields?
    - Or, or only use timestamp, units, and values due to potential inaccuracies of tags
- **Identify if any feature engineering can be done to aid classification**
    - Research how to implement feature engineering
    - Brainstorming features
    - Decide what features to create
    - Create those features (maybe move to model phase)
    - Decide what features to import
        - Ex. Pulling in more detailed data from the csvs Jiachen showed us, pulling in data from the ION database, etc...
    - Check how features work with our model (maybe move to model phase)