Weekly Sprint Planning

2020-05-22 /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: Modeling and feature selection

2. Which tasks did we complete?
   - **Research approach to populating missing data**
   - **Develop the code to identify when fields are null**
   - **Modify current project objectives to a new report**
   - **Confirm details for project objective**
   - **Identify relevant metadata fields**
   - **Preliminary data labeling**
3. Was there anything stopping us from finishing specific tasks?
   - Lack of clarity on sensor tags (stopped us from finishing data labeling)
   - Database design issues (field data type all as strings)
     - Blocks data aggregation

4. What tasks are still in progress?
    - Identify relevant features
    - Transform data for ML tasks
    - Make training and testing data
    - Identify if any feature engineering can be done

5. Are there any changes that need to be made?
    - Update flow charts when appropriate
6. What is our goal/theme for this week?
   Goal: Feature engineering and modeling

7. What tasks need to be added/replenished to the Backlog?

8. What tasks are most important and should be pulled from Backlog to In progress?
    - Task 1
    - Task 2
    - Task 3

9. Are there any dependencies between In Progress tasks?
    a. If so, how will that be organized?

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

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

- **Making training and testing data (Alex & Claudia)**
    - Finalize labeling
- **Feature Selection/Feature Engineering**
    - Update flowchart (Eva)
    - Make categorical fields into smaller levels (Claudia & Eva)
    - Code to calculate sensor update rate (Alex)
    - Code to aggregate numeric values (Connor)
    - Code to aggregate boolean values
    - Code to aggregate string (state) values?
- **Develop various models for NRCan tags (Connor)**
    - Identify methods to incorporate how NRCan sensors vary with respect to the non-NRCan sensor values
        - Cluster non-NRCan sensors and do the following per group and use these values as predictors for NRCan labels
            - Average non-NRCan sensor value per hour of the day
            - Average daily non-NRCan sensor value
            - Average non-NRCan sensor value per month (or week?)
        - Cluster non-NRCan sensors and fit a linear model to each NRCan sensor with the clusters as predictor variables, take the coefficients from the model fits and use as predictor variables for the NRCan sensors
    - Supervised (possible options that may work well with our data, b/c quite flexible)
        - Random forest
        - Neural Net
            - ANN or RNN?
        - Bagging
        - Boosting
    - Semi-Supervised (possible options that may work well with our data)
        - Label Spreading

        - Label Propagation
        - HiddenMarkovModel Semi-supervised
        - BayesClassifier Semi-supervised
        - NaiveBayes Semi-supervised
    - Unsupervised (not sure if this will work but it may be worth looking into if the others don't? Probably more of an interesting thing to try after the capstone is done as it may take too long to implement)
        - Boltzmann Machine (an unsupervised Neural Net)
            - Boltzmann Machines attempt to model a system and interpret how different sensors interact by taking values and applying weight to nodes (commonly used in recommendation systems)
            - Theoretically we could create a Boltzmann Machine to emulate the system with the different NRCan tags as the visible nodes and then calculate the probability of a given sensor belonging to that node given the current (or average) reading from the sensor
- **Develop various models for non-NRCan tags (Connor)**
    - Decide which models to implement (meet and brainstorm?)
        - Different for NRCan tag-able and non-NRCan tag-able sensors
    - DBSCAN (Accounts for noise values so they don't mess with the clusters)
    - HDBSCAN (Accounts for noise values so they don't mess with the clusters)
    - Gaussian Mixture Models
    - Variational Bayesian estimation of a Gaussian mixture
    - K-Means
    - Hierarchical
    - Fuzzy C-Means
    - Mean Shift
- **Compare models for effectiveness (move into Scaling?)**

    - Research ways to evaluate models/choose performance metrics for comparison
    - Develop function for the chosen performance metrics (if needed)
    - Outline reasons for chosen models
    - Develop code to grid search for optimal model/model configuration/features
- **Recommend models (move into Scaling?)**
    - Create a flowchart/report to show chosen model and outputs