Week 4 Report

Summary

Full models results:

Rd_particle_daily: R-adj: 0.716, Deviance=80%

Rd_particle_hourly: R-adj: 0.731, Deviance = 73.6%

Radon_daily: R-adj: 0.716, Deviance = 80.2%

Radon_hourly: R-adj: 0.631 Deviance = 63.7%

- the full models needs to be truncated, but the method of removing variables is still to be investigated as of right now

Regsubset Results:

Common variables identified by all three criteria: listed in red, see below

Common variables identified by all three criteria and appears in both Radon and Rd-particle: listed in Green

- **wsp_mean, co2_ppm_mean, ch4_mean, h2o_sync_mean, ethene_mean**
  - For daily (it means that these variables were common in all of adj r-squared, cp, bic selections when fitted on both radon_daily and rd-particle_daily)
- datetime, no2, temp_f, pressure_altcorr, wsp, wdr, relh, co2_ppm, ethane, propene, X1_3.butadiene, i.butane, n.butane, acetylene, benzene, hour
  - for hourly

Regsubset on Radon:

Daily

1. using Adj_R^2: 21 variables

    a. Intercept, o3_mean, **temp_f_mean, pressure_altcorr_mean, wsp_mean**, wdr_mean, rain_mean, **co2_ppm_mean, ch4_mean, h2o_sync_mean, ethene_mean**, i.butane_mean, n.butane_mean, **acetylene_mean, cyclopentane_mean, i.pentane_mean, n.pentane_mean**, n.hexane_mean, benzene_mean, ethyl.benzene_mean, m.p.xylene_mean, o.xylene_mean

2. Using Cp, 14 variables

    a. Intercept, **temp_f_mean, pressure_altcorr_mean, wsp_mean**, wdr_mean, rain_mean, **co2_ppm_mean, ch4_mean, h2o_sync_mean, ethene_mean, acetylene_mean, cyclopentane_mean, i.pentane_mean, n.pentane_mean**, benzene_mean

3. Using Bic, 11 variables

    a. Intercept, **temp_f_mean, pressure_altcorr_mean, wsp_mean**, **co2_ppm_mean, ch4_mean, h2o_sync_mean, ethene_mean**, **acetylene_mean, cyclopentane_mean, i.pentane_mean, n.pentane_mean**

Hourly

1. using Adj_R^2: 25 variables

    a. Intercept, datetime, co, no2, nox, temp_f, pressure_altcorr, wsp, wdr, relh, co2_ppm, h2o_sync, ethane, propene, X1_3.butadiene, i.butane, n.butane, acetylene, cyclopentane, n.pentane, n.hexane, benzene, ethyl.benzene, m.p.xylene, o.xylene, hour

2. Using Cp, 20 variables

    a. Intercept, datetime, no2, nox, temp_f, pressure_altcorr, wsp, wdr, relh, co2_ppm, ethane, propene, X1_3.butadiene, i.butane, n.butane, acetylene, cyclopentane, n.hexane, n.heptane, benzene, hour

3. Using bic, 16 variables

    a. Intercept, datetime, no2, temp_f, pressure_altcorr, wsp, wdr, relh, co2_ppm, ethane, propene, X1_3.butadiene, i.butane, n.butane, acetylene, benzene, hour

Regsubset on Rd-particle:

Daily:

1. Using adj-rsqaured: 15 variables
   a. Intercept, co_mean, o3_mean, wsp_mean, co2_ppm_mean, ch4_mean, h2o_sync_mean, ethene_mean, cyclopentane_mean, i.pentane_mean, n.pentane_mean, isoprene_mean, n.octane_mean, toluene_mean, m.p.xylene_mean, o.xylene_mean
2. Using cp: 9 variables
   a. Intercept, o3_mean, wsp_mean, co2_ppm_mean, ch4_mean, h2o_sync_mean, ethene_mean, cyclopentane_mean, n.hexane_mean, toluene_mean
3. Using BIC: 8 variables
   a. Intercept, o3_mean, wsp_mean, co2_ppm_mean, ch4_mean, h2o_sync_mean, ethene_mean, n.heptane_mean, toluene_mean
4. Comparing with Radon
   - Intercept, **temp_f_mean, pressure_altcorr_mean, wsp_mean**, **co2_ppm_mean, ch4_mean, h2o_sync_mean, ethene_mean**, **acetylene_mean, cyclopentane_mean, i.pentane_mean, n.pentane_mean**
   - **wsp_mean, co2_ppm_mean, ch4_mean, h2o_sync_mean, ethene_mean,**
     - overlapping ones from all 6 criteria

Hourly:

1. Using adj-rsqaured: 25 variables
   a. Intercept, datetime, co, no2, nox, temp_f, pressure_altcorr, wsp, wdr, relh, co2_ppm, h2o_sync, ethane, propene, X1_3.butadiene, i.butane, n.butane, acetylene, cyclopentane, n.pentane, n.hexane, benzene, ethyl.benzene, m.p.xylene, o.xylene, hour
2. Using cp: 20 variables

    a. Intercept, datetime, no2, nox, temp_f, pressure_altcorr, wsp, wdr, relh, co2_ppm, ethane, propene, X1_3.butadiene, i.butane, n.butane, acetylene, cyclopentane, n.hexane, n.heptane, benzene, hour

3. Using BIC: 16 variables
    a. Intercept, datetime, no2, temp_f, pressure_altcorr, wsp, wdr, relh, co2_ppm, ethane, propene, X1_3.butadiene, i.butane, n.butane, acetylene, benzene, hour
4. Comparing with Radon
    - Intercept, datetime, no2, temp_f, pressure_altcorr, wsp, wdr, relh, co2_ppm, ethane, propene, X1_3.butadiene, i.butane, n.butane, acetylene, benzene, hour
    - datetime, no2, temp_f, pressure_altcorr, wsp, wdr, relh, co2_ppm, ethane, propene, X1_3.butadiene, i.butane, n.butane, acetylene, benzene, hour
        - literally the same with radon

Small update on find highly-correlated variables and remove them by using only the ones that highly correlate with the response

- weighted count was indeed better than count so the algorithm was correct
- problem was some predictors were just not as correlated with the response but were included because they don’t really correlated with any other predictors
    - for example **neither weighted.count nor count should have been included, but weighted.count remained since it was better than count, rain** is another example which **doesn’t correlate with response but since it doesn’t correlate with any other predictors,** it was not eliminated
- Thus algorithm needs to be improved

To do for Week 4:

- do similar for rd-particle

- find highly-correlated variables and remove them by using only the ones that highly correlate with the response
- do a pairwise correlation between all of the predictors and response for the fitted model
- XG-boost
- *do 50 km for flares, and do count and closest flare, and if possible do a weighted count*

Task 1: do 50 km for flares, and do count and closest flare, and if possible do a weighted count

- changed the boundary to a square 100 km from Loving's site
    - more distance to the west
        - didn't change much
    - calculated a weighted count and added closest distance
        - weighted count = summation of (1 / distance)
        - closest = 100 if no flare on that day within 100km (count = 0, weighted_count = 0)
    - the new variables (weighted count and closest distance) didn't contribute much

Task 2: do a pairwise correlation between all of the predictors and response for the fitted model

- cor_matrix, changed the labels to be on the side

Task 3: find highly-correlated variables and remove them by using only the ones that highly correlate with the response

- threshold higher ⇒ more variables, more collinearity
- problem with the order of removing the predictors

    - count vs weighted.count, even though count is higher correlated with the response, it was probably removed earlier when compared to another variable
    - but weighted.count was untouched since it didn't correlate with any other variable than count, it remained, leading to choosing a suboptimal variable

Task 4: do similar for rd-particle

Normal Gam_daily: R-adj: 0.716, Deviance=80%

Normal Gam_hourly: R-adj: 0.731, Deviance = 73.6%

Recall for Radon:

- daily: 0.716 and 80.2%
- hourly: 0.631 and 63.7%

- Rd-particle performs better on hourly

Regsubset results

Daily:

1. Using adj-rsqaured: 15 variables
    a. Intercept, co_mean, o3_mean, wsp_mean, co2_ppm_mean, ch4_mean, h2o_sync_mean, ethene_mean, cyclopentane_mean, i.pentane_mean, n.pentane_mean, isoprene_mean, n.octane_mean, toluene_mean, m.p.xylene_mean, o.xylene_mean
2. Using cp: 9 variables
    a. Intercept, o3_mean, wsp_mean, co2_ppm_mean, ch4_mean, h2o_sync_mean, ethene_mean, cyclopentane_mean, n.hexane_mean, toluene_mean
3. Using BIC: 8 variables
    a. Intercept, o3_mean, wsp_mean, co2_ppm_mean, ch4_mean, h2o_sync_mean, ethene_mean, n.heptane_mean, toluene_mean
4. Comparing with Radon
    - Intercept, **temp_f_mean, pressure_altcorr_mean, wsp_mean**, **co2_ppm_mean, ch4_mean, h2o_sync_mean, ethene_mean**,

acetylene_mean, cyclopentane_mean, i.pentane_mean, n.pentane_mean

- **wsp_mean, co2_ppm_mean, ch4_mean, h2o_sync_mean, ethene_mean,**
  - overlapping ones from all 6 criteria

Hourly:

1. Using adj-rsqaured: 25 variables
   a. Intercept, datetime, co, no2, nox, temp_f, pressure_altcorr, wsp, wdr, relh, co2_ppm, h2o_sync, ethane, propene, X1_3.butadiene, i.butane, n.butane, acetylene, cyclopentane, n.pentane, n.hexane, benzene, ethyl.benzene, m.p.xylene, o.xylene, hour
2. Using cp: 20 variables
   a. Intercept, datetime, no2, nox, temp_f, pressure_altcorr, wsp, wdr, relh, co2_ppm, ethane, propene, X1_3.butadiene, i.butane, n.butane, acetylene, cyclopentane, n.hexane, n.heptane, benzene, hour
3. Using BIC: 16 variables
   a. Intercept, datetime, no2, temp_f, pressure_altcorr, wsp, wdr, relh, co2_ppm, ethane, propene, X1_3.butadiene, i.butane, n.butane, acetylene, benzene, hour
4. Comparing with Radon
   - Intercept, datetime, no2, temp_f, pressure_altcorr, wsp, wdr, relh, co2_ppm, ethane, propene, X1_3.butadiene, i.butane, n.butane, acetylene, benzene, hour
   - datetime, no2, temp_f, pressure_altcorr, wsp, wdr, relh, co2_ppm, ethane, propene, X1_3.butadiene, i.butane, n.butane, acetylene, benzene, hour
     - literally the same with radon

Task 5: XG-Boost