HSPH Capstone Project Milestone 3

Members: Casey, Chris, Justin, and Keyan

TF: David Sondak

Big Picture

- Completed imputation and modeling on 1% subset of the data
- Built CNN architecture and have written imputation scripts that are ready to be used on the full data
- We should obtain preliminary full-scale modeling results for the milestone

Progress - Infrastructure

- Data preprocessing/imputations done on Odyssey
  - Combination of R and Python
  - Census data incorporation, dropping redundant and unimportant variables, feature engineering

- Successfully installed PyTorch on Odyssey and have trained basic models

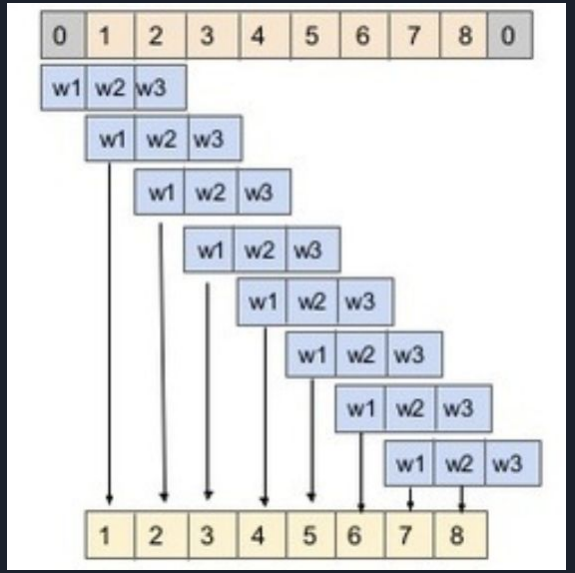
Progress - Imputations

- Previously ran into memory issues using the R package “missForest”

- Attempting to use a Python implementation of the missForest algorithm called “predictive_imputer”
  - Can use Python’s “pickle” package to save a model trained on a subset, for later imputation on the full data set

CNN Architecture

- Use 1D temporal convolutions for non-static variables to make use of days for which there are no PM2.5 outputs
- Merge static variables (e.g. census) with hidden outputs from convolutions and use feed-forward NN from there



Other Models

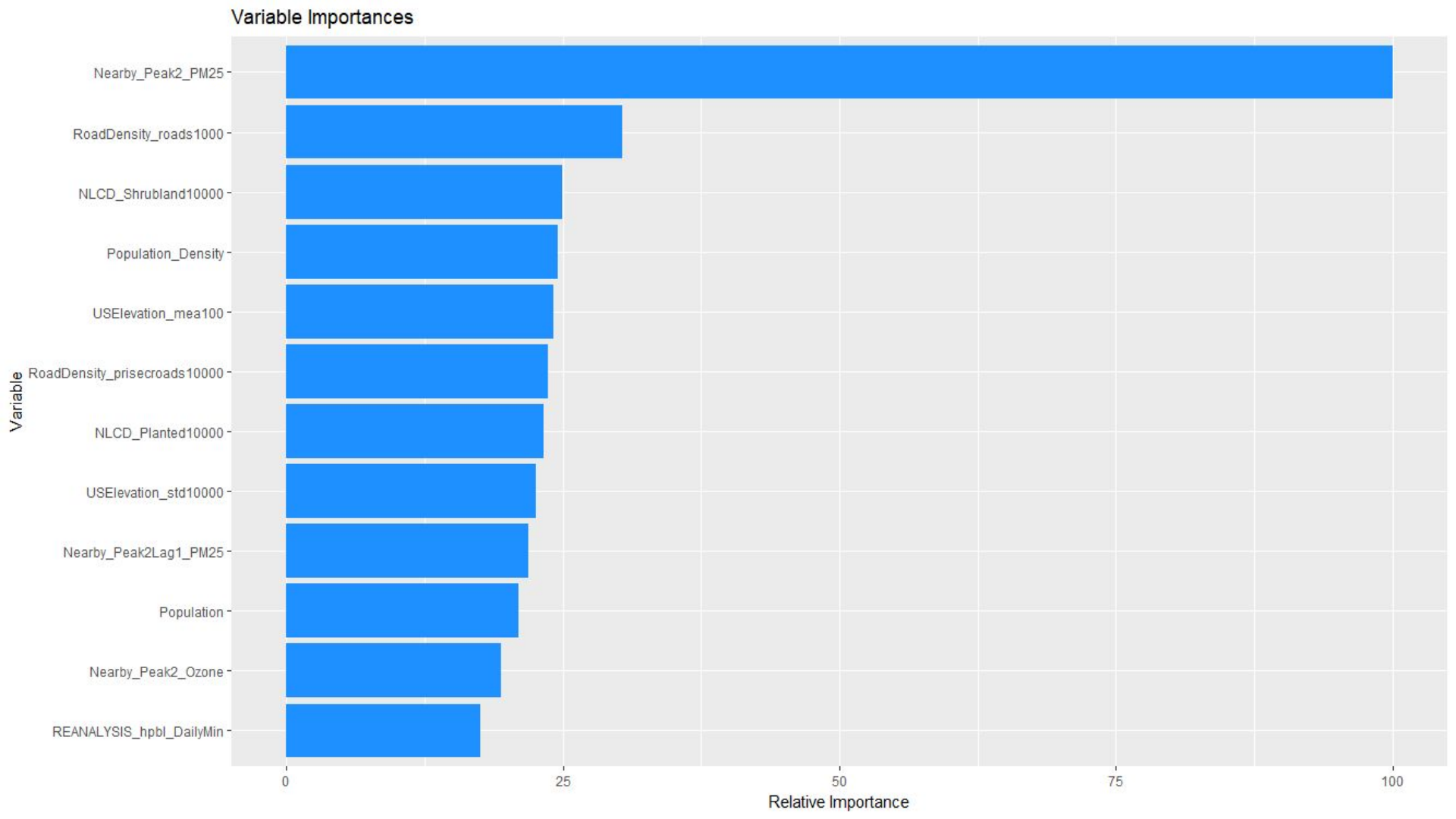
- Use scikit-learn models such as random forest, ridge/lasso regression, Gaussian process regression on full data for comparison

- These models don’t make use of days for which there are no PM2.5 outputs

- Easy to use and test

Hurdles - Variable Importance

- Nearby $PM_{2.5}$ levels have disproportionately high predictive power on $PM_{2.5}$ levels
    - OLS model of $PM_{2.5}$ vs. Nearby $PM_{2.5}$ - R^2 of 0.75
    - Addition of other predictors with more complex models add only small improvements - need a sense for how much improvement is possible
    - Need to discuss with HSPH team how this issue was addressed previously


Variable Importances
Nearby_Peak2_PM25
RoadDensity_roads1000
NLCD_Shrubland10000
Population_Density
USElevation_mea100
RoadDensity_prisecroads10000
NLCD_Planted10000
USElevation_std10000
Nearby_Peak2Lag1_PM25
Population
Nearby_Peak2_Ozone
REANALYSIS_hpbl_DailyMin
Variable
0
25
50
75
100
Relative Importance

Future Steps

- Apply CNN architecture to full dataset, compare R^2
- Build infrastructure for adding more static/dynamic variables
- Discuss framework with HSPH team; incorporate their feedback
- Documentation!