Homework 3
Yanran Li (yl5465)
November 21, 2023

1 Introduction

Heat warnings are issued in advance of forecast extreme heat events, yet little evidence is available regarding their effectiveness in reducing heat-related illness and death[1]. In this homework 3, I implemented a Bayesian Additive Regression Trees model (an ensemble method) on a real-world data set (heat warning related) to estimate the association between heat index and other variables ($PM_{2.5}$ etc.).

2 Data

- **Description of the dataset:** I used the daily time series data[2], acquired during the warm months (April-October) of 2006-2016 for 2837 U.S. counties. For each county, we obtained 1) daily maximum heat index (an index that combines air temperature and relative humidity to posit a human-perceived equivalent temperature); 2) daily issuance of heat alerts (binary); 3) daily estimated concentrations of fine particulate matter (with an aerodynamic diameter of less than 2.5 μm; $PM_{2.5}$).

- **Description of the response variable:** *Heat Index:* We obtained gridded (4-km resolution) estimates of daily maximum temperature and vapor-pressure deficit for the contiguous US from the Parameter-elevation Regressions on Independent Slopes Model. From these variables, we generated time series of population-weighted daily maximum heat index for each county[1]. The visualization of the "Heat Index" variable over time has already been shown.

- **Description of the features:** In hw 1, we've already shown the distribution of all features we have as well as confounders (distributions also shown in Figure 5 in appendix of this hw). Specifically in this homework, I focus on the Heat Index in 2007. We could model the Heat Index ("HImaxF_PopW") to discover its relationship with other features: we have chosen four covariates: annual average $PM_{2.5}$, county's smoke rate, population size of that county, and whether is a weekday or a weekend.

- **Number of observations:** After processed null value, I obtain a feature matrix X with 442338×4 elements and a observation matrix Y with dimension 442338×1.

3 Methods

3.1 Probabilistic Model

Bayesian additive regression trees (BART) is a non-parametric regression approach. If we have some covariates X, and we want to use them to model Y, a BART model can be represented as:

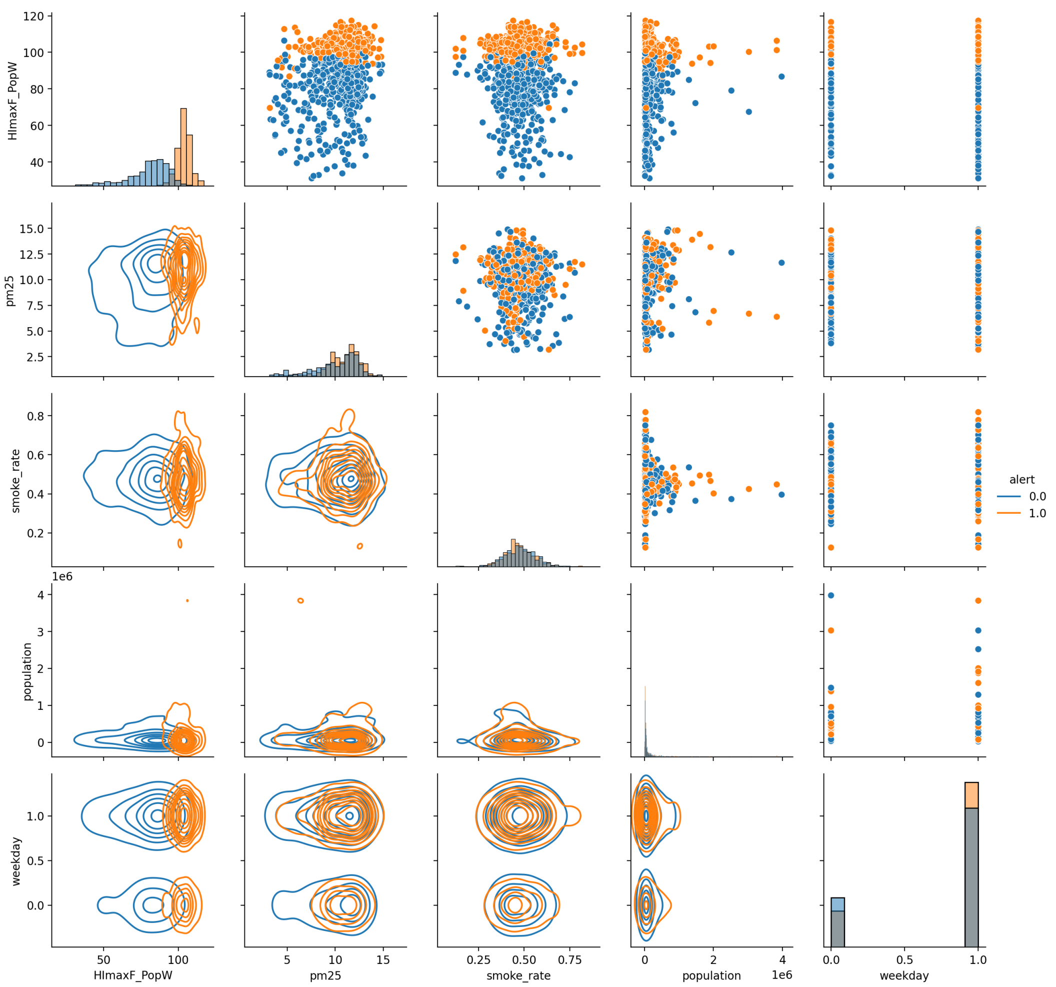

Figure 1: Visualize the relationship between different features

$Y = f(X) + \epsilon$, where we use a sum of m regression trees to model f, and ϵ refers to the noise. A key idea is that a single BART-tree is not very good at fitting the data but when we sum many of these trees we get a good and flexible approximation.

Given Y, we assume that each Y_i could be represented by 2 latent components (α and μ). The data generating process is:

1. Exponential Distribution for α:

$$\alpha \sim \text{Exponential}(\lambda = 1)$$

2. BART Model for μ:

$$\mu = \text{BART}(X, \log(Y), m = 50)$$

The BART model is a sum-of-trees ensemble model. It uses X to predict $\log(Y)$ (the log-transformed value of "HImaxF_PopW"). The hyperparameter m denotes we used 50 trees in our BART Model for μ.

3. Negative Binomial Distribution for Y:

$$Y \sim \text{NegativeBinomial}(\exp(\mu), \alpha)$$

Y is assumed to follow a Negative Binomial distribution, where the expected mean is modeled as the exponential of the output from the BART model. α serves as the dispersion parameter.

3.2 Inference Method

The model parameters are estimated using MCMC sampling, which approximates the posterior distribution of the model parameters given the observed data. From the implementation side, we use the The "pm.sample()" function from PyMC library, and utilize MCMC to draw samples from the posterior distributions of α and the parameters underlying the BART model, providing a Bayesian inference for the expected Heat Index ("HImaxF_PopW").

4 Result

4.1 Convergency

For the non-BART variables α, I plot its trace plot (Figure 2) by using the `arviz.plot_trace` function.

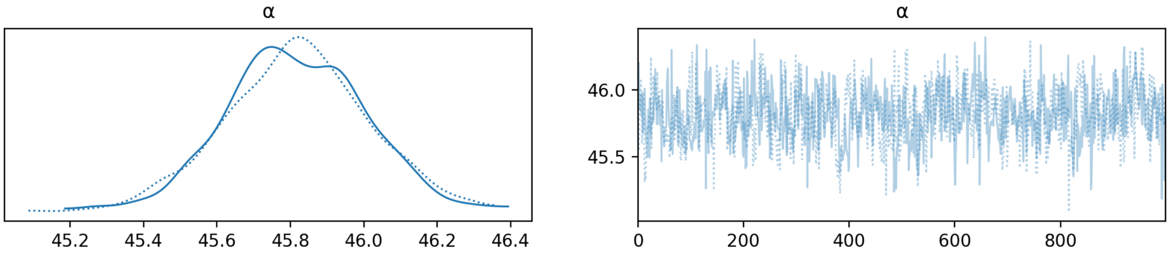

Figure 2: Trace plot for latent variable α

For the BART variables (μ) I used `pymc_bart.plot_convergence` function to check its R-hat (<= 1.01), and Effective Sample Size (ESS) numerical diagnostics (Figure 3). The blue line is the empirical cumulative distribution for those values. For the ESS we want the entire curve above the dashed line, and for R-hat we want the curve to be entirely below the dashed line. In Figure 3, we can see that we make it for the ESS and for the R-hat.

4.2 Evaluation

To evaluate the goodness of fit of the model, we used partial dependence plots (Figure 4) to show the marginal effect that one covariate has on the predicted variable. That is, what is the effect that a covariate X_i has of Y while we average over all the other covariates ($X_j, \forall j \neq i$).

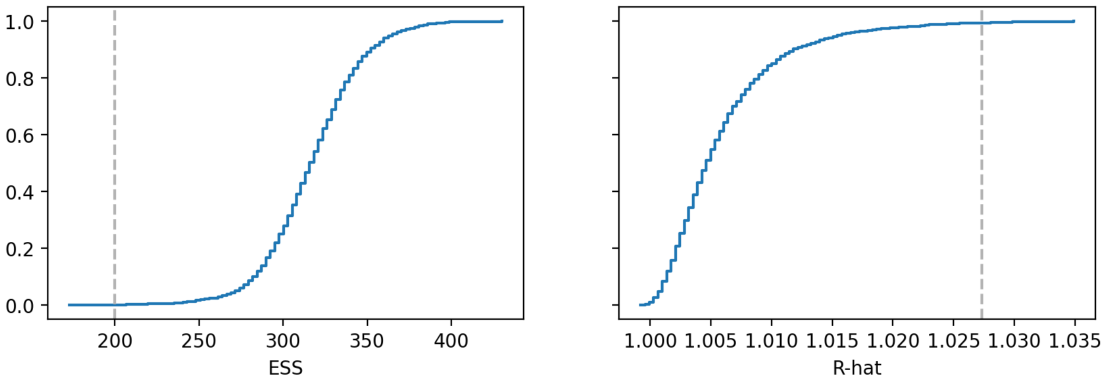

Figure 3: Numerical diagnostics for latent variable μ

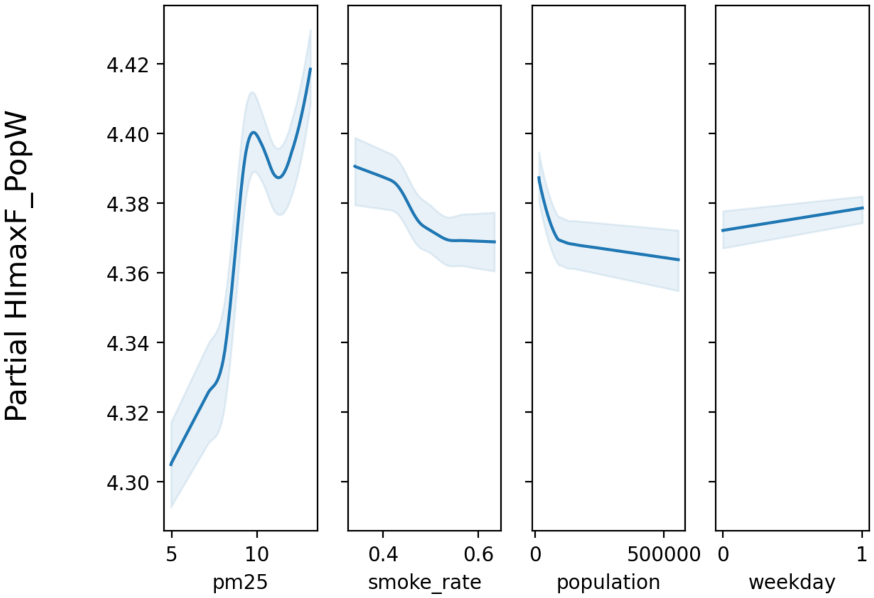

Figure 4: Partial dependence plots

From figure 4 we can see the main effect of each covariate on the predicted value. For the $PM_{2.5}$ covariate, we can see around 10 it will have a peak in effect for Heat Index ("HImaxF_PopW"), but decrease after that before its next increasing. We also plot the relative importance in a scale from 0 to 1 (less to more importance) and the sum of the individual importance is 1 (Appendix figure 6), where the relative importance qualitative agrees with the partial dependence plot. The covariate "$PM_{2.5}$" has the highest importance and the "weekday" has the lowest importance in our model.

References

[1] Kate R. Weinberger, Xiao Wu, Shengzhi Sun, Keith R. Spangler, Amruta Nori-Sarma, Joel Schwartz, Weeberb Requia, Benjamin M. Sabath, Danielle Braun, Antonella Zanobetti, Francesca Dominici, and Gregory A. Wellenius. Heat warnings, mortality, and hospital admissions among older adults in the united states. *Environment International*, 157: 106834, 2021. ISSN 0160-4120. doi: https://doi.org/10.1016/j.envint.2021.106834. URL https://www.sciencedirect.com/science/article/pii/S0160412021004591.

[2] Xiao Wu, Kate R Weinberger, Gregory A Wellenius, Francesca Dominici, and Danielle Braun. Assessing the causal effects of a stochastic intervention in time series data: are heat alerts effective in preventing deaths and hospitalizations? *Biostatistics*, page kxad002, 02 2023. ISSN 1465-4644. doi: 10.1093/biostatistics/kxad002. URL https://doi.org/10.1093/biostatistics/kxad002.

5 Appendix

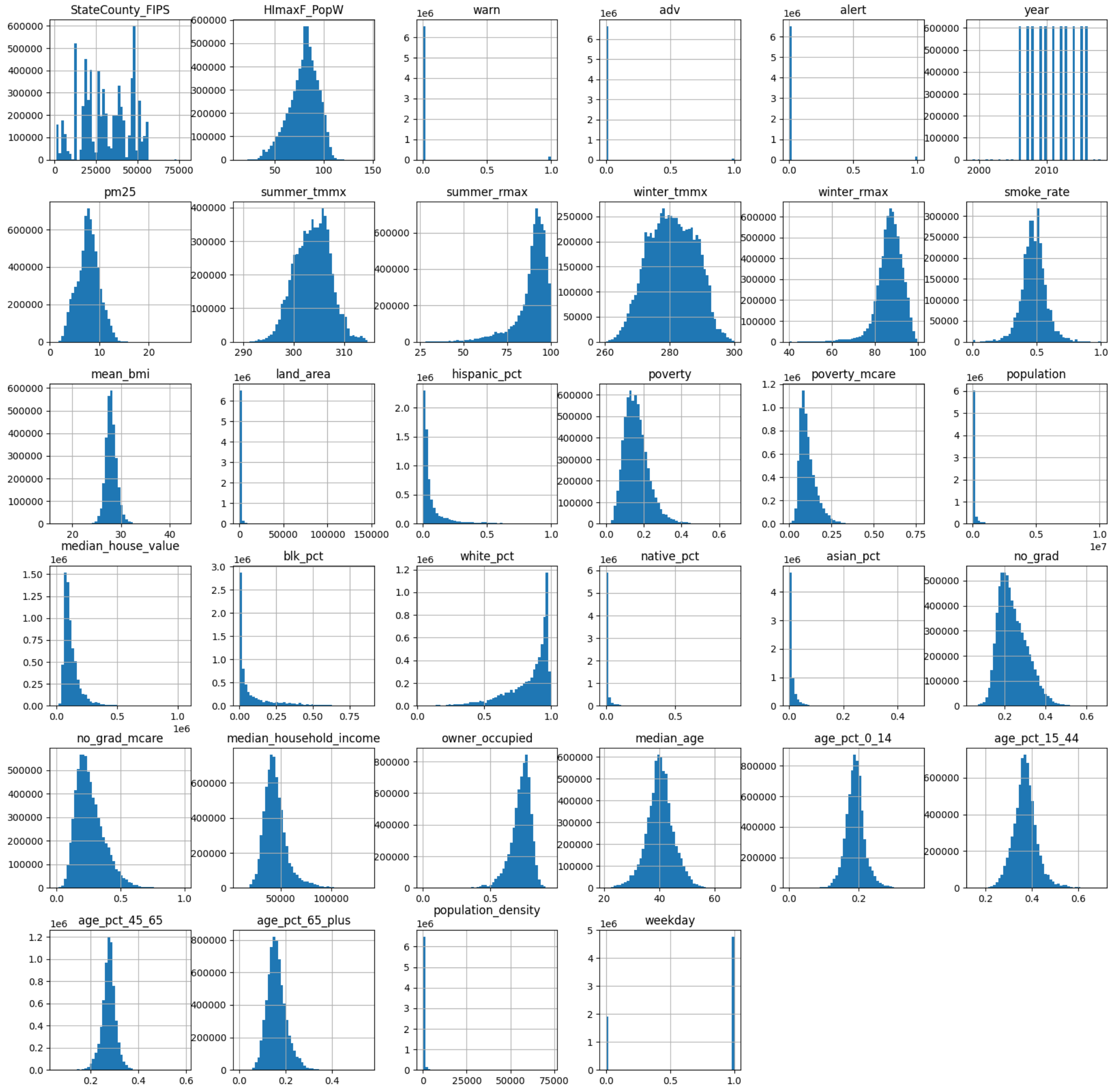

Figure 5: Histogram of all features

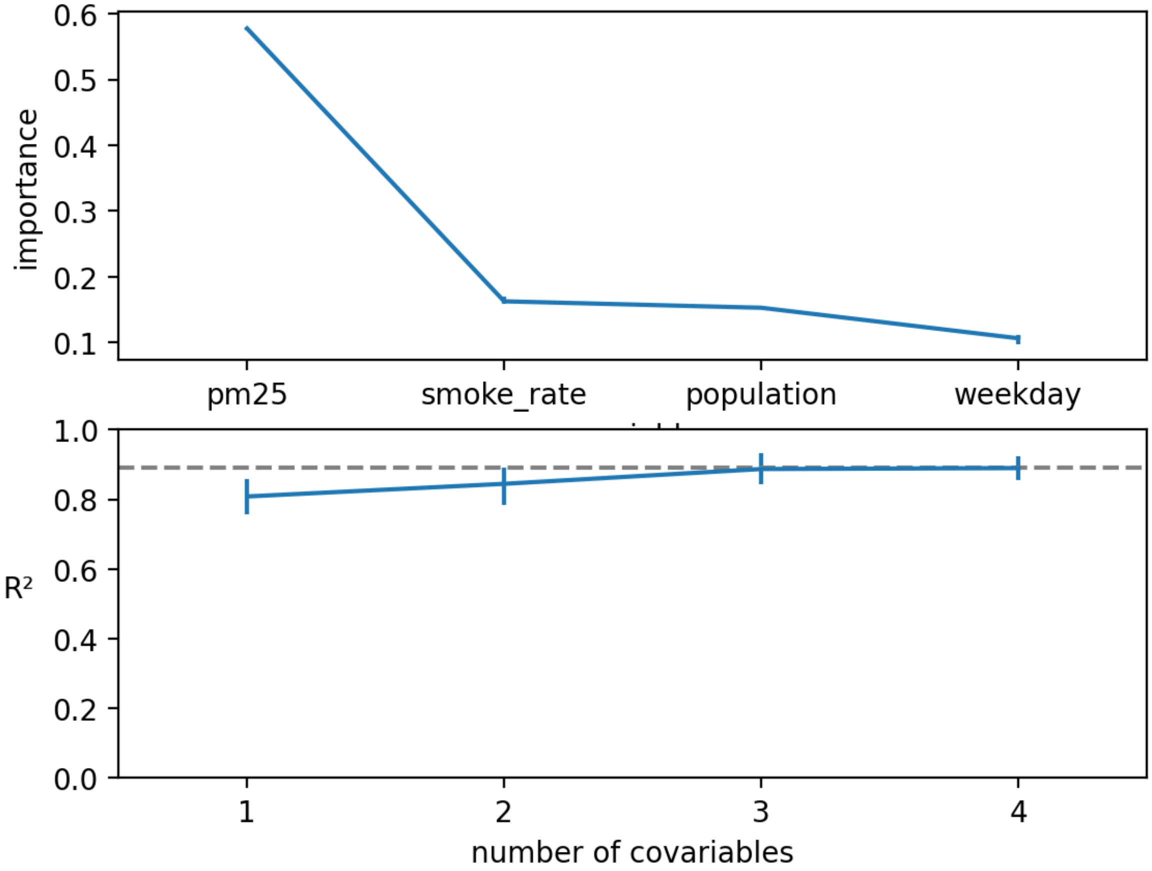

Figure 6: Variable importance