A Bayesian Approach to Smoke Detection

Julia Burek, Gunnar Franko, Chelsea Le Sage

Problem Description

The successful detection of smoke in the presence of fire or similar emergencies is a public safety concern. In the United States, smoke alarms are present in 96% of homes, but 20% of these alarms are non-operational ("Smoke Alarm Research"). Environmental factors such as humidity, temperature, and air contamination influence the rate of smoke detection, further listed in Appendix J. False fire alarms (type I error) are also a concern for first responders. False alarms could be triggered by cooking fumes, aerosol sprays, dust, and more ("8 Reasons Why…"). Our goal for this project is to have a better understanding of these fire-related variables for informed safety practices and explicit fire identification. We would like to create a logistic regression model to predict the presence of fire as well as determine the variables that are most informative for this detection. We will use both Bayesian sampling and variational inference methods to approach this problem. This model is important as AI or machine learning assisted devices could be the future of smoke alarms. Metadata could feasibly be added to smoke sensors to improve detection accuracy and public safety.

Probability Model

The first model that we used was a main effects logistic regression model. The graphical representation of this model is shown in Appendix E. We used Gaussian priors for the intercept and the predictor variables that had uninformed hyperparameters. The posterior distribution was Bernoulli: 0 for no fire present and 1 for fire present. This model included all of the prediction variables.

After performing some initial exploratory data analysis, we went forward with creating the logistic regression model. We used a full model, including all of the predictor variables for the HMC approach. We wanted to see if there was a difference between the full model and a reduced model with some predictors dropped. We used the ADVI approach for both the full model as well as the reduced model. The main effects logistic regression model from a Bayesian approach results in the following equation:

$$Pr(Y = 1 \mid X = x) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

For the input $(x_i, y_i)_{i=1}^{n}$ assuming that $x_i = (x_{i0}, \ldots, x_{ip})$ and $y_i \in \{0, 1\}$ for the Bernoulli function $y_i|x_i \sim Bern(p_i(\beta_0 + \beta_1))$.

Approach

The data cleaning performed was minimal as our dataset is robust with no missing values. We dropped columns that indexed the observations as well as the timestamp column. From here we performed exploratory data analysis in order to understand the interactions between our predictor variables before building any models. Appendix A reveals that the distribution of our target variable 'Fire Alarm' is not equal; for 71% of observations fire was present (truth value 1) while the other 29% indicate an absence of fire (truth value 0). Ideally this target variable would not be imbalanced, rather we would see a 50-50 distribution.

The correlation coefficients between predictors and the response is shown in Appendix B. We find that humidity has the strongest correlation with the target variable, with a value of 0.4. This can also be observed in the heat map (Appendix C). Moreover, the heat map reveals a correlation above 0.9 between particulate matter (PM) variables and number concentration of particulate matter (NC) variables. Such correlation coefficients could indicate that the predictors are redundant and could be dropped to minimize model complexity. This is why we wanted to attempt a reduced model.

In order to ensure that there was a difference in the data when comparing observations in the presence of fire versus observations without fire, we performed a Bayesian analysis of the posterior marginal probability distributions to explore the posterior means of the fire and no fire groups. This started by a standard scaling of the data and then splitting it between the two target response groups. From there we calculated the posterior probability for each variable as shown in Appendix D. Looking at these results we found that for most variables there is a significant difference between the two means. For the variables that we do not find much of a difference, we discover later in our analysis that our model could do without them and perhaps perform better without them.

We approached our logistic regression model by starting with a Hamiltonian Monte Carlo (HMC) sampling approach. We quickly found that sampling was not well-suited for the 60,000 observations in our data. Instead, we randomly selected 10% of the data to run our HMC main effects model. We chose the HMC model because all of our predictor variables were continuous. It also moves to interesting regions faster, and we were not concerned about the increased cost of computation because we were only looking at a fraction of our data. We generated trace plots from the sampling to check for convergence, then used our model to perform prediction and check the accuracy of our model, discussed further in the next section.

Due to the size of the dataset and its multivariate nature, we turned to variational approximation. To do this we used an ADVI model using the entire dataset with 10,000 samples. With this model we looked at the ELBO plot to check for convergence as well as forest plots to observe the certainty of the predictor variables. After seeing convergence, we used the ADVI model to predict the response variable for our samples and checked the accuracy of this model, discussed further in the next section.

Results

After performing an analysis of the posterior marginal probability distributions to explore the posterior means of the fire (1) versus the no fire (0) groups, we found that there was a significant difference between the posterior means of each group. This was supported by running an F-test on the difference of Gaussian means with a null hypothesis of the mean vector for both groups being equal and an alternative hypothesis that the mean vectors are different. The test resulted in a T-value of 2.2E+4. As this value is so large, we can reject the null hypothesis that the mean vectors of the groups are the same. Understanding that there is a significant difference between the fire and no fire groups, we moved forward with our logistic regression model using a Bayesian approach.

We first ran the Hamiltonian Monte Carlo main effects logistic regression model on a random sample of 10% of the data. We included all of the predictor variables. After creating the model using 1,000 samples, we created trace plots and density plots from the sampling to check for convergence. We determined there was convergence of most of the variables, and that their distributions appeared normal. We then checked for accuracy of the model in predicting the posterior distribution of the presence of fire. The accuracy was around .90 for the samples. This indicated the model did a good job at predicting the presence of fire given the using all of the predictor variables. Because the HMC sampling approach could only work efficiently on 10% of the data, we wanted to move forward with a variational inference approach using ADVI on the full dataset to create a main effects logistic regression model.

Due to the size of the dataset and its multivariate nature, we turned to variational approximation. To do this we used an ADVI model with the entire dataset with 10,000 samples. After creating this model with 10,000 samples, we looked at the ELBO plot to check for convergence as well as forest plots to look at the uncertainty among the predictor variables. After seeing convergence in the ELBO plot, we used the ADVI model to predict the presence of fire and check the accuracy of this model. The accuracy of the model was approximately 0.80. The forest plot revealed that 6 of the predictor variables had confidence intervals that straddled 0, indicating a higher degree of uncertainty in the eCO2, PM, and NC predictors. Due to the aforementioned high correlation between all PM and NC variables (Appendix C), this behavior was not surprising. Thus for the reduced model we dropped eCO2 and PM1.0 as predictors to see if this uncertainty influenced the accuracy of our model. After running the reduced model, we found that the model had an accuracy of approximately 0.79. This was not a significant difference from the full model. In order to include more predictors, we would choose the full main effects logistic regression model rather than the reduced one.

The results were on par to what we would expect using both the HMC sampling approach as well as the variational inference approach. From our previous knowledge, sampling typically yields a more accurate prediction while variational inference can only provide an approximation. This was similar to our results. Ultimately, we would want to choose the ADVI approach given its efficiency and ability to work with the large dataset and its multivariate nature. Our model's

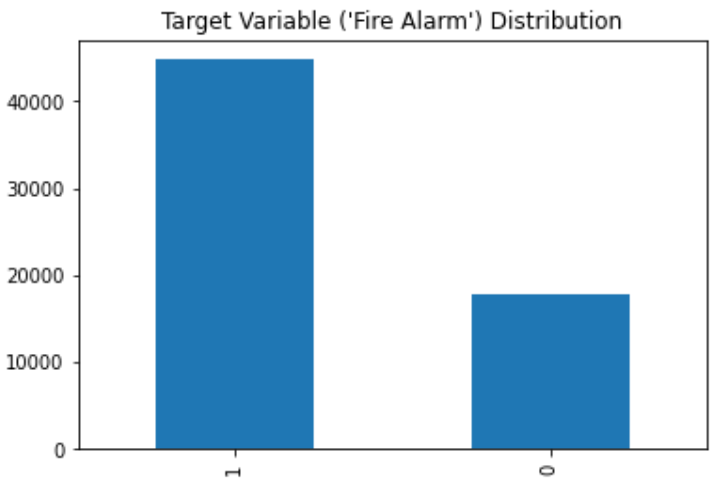
relatively high accuracy indicates strong performance; however, given the imbalanced nature of the data, the model should be accepted with caution.

Conclusions

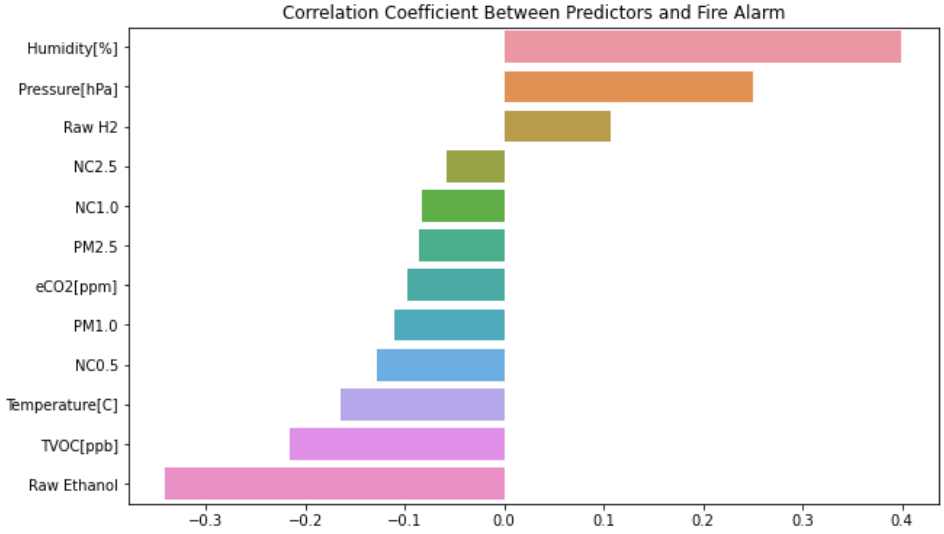
The objective of our research was to determine the major predictors of fire given smoke detection data. We wanted to find an efficient and accurate logistic regression model to predict the presence of a fire using Bayesian approaches, sampling and variational inference. As expected, we found high accuracy with our sampling approach but, due to processing capabilities and the size of our data, only used 10% of our available observations to create the model. Variational approximation, however, allowed us to use every observation but produced less accurate results than the HMC sampling approach. We did see convergence in the ELBO plot, however. Overall, our findings indicate that machine learning models could be utilized in smoke sensors to improve the accuracy of fire detection and to lower the rate of false alarms. It is important to acknowledge that the models we created are not necessarily the best models for this problem. It is also important to acknowledge the issue of imbalanced data. It could be beneficial to counteract this imbalance somehow in order to develop a more accurate prediction model. Additionally, our work could be more impactful with additional information regarding whether a smoke alarm triggered or not. This way we would be able to generate a confusion matrix to analyze the issue of false alarms given the environmental predictors. Overall, this project showed how Bayesian Machine Learning can be helpful when solving an important problem such as fire detection.

Appendices

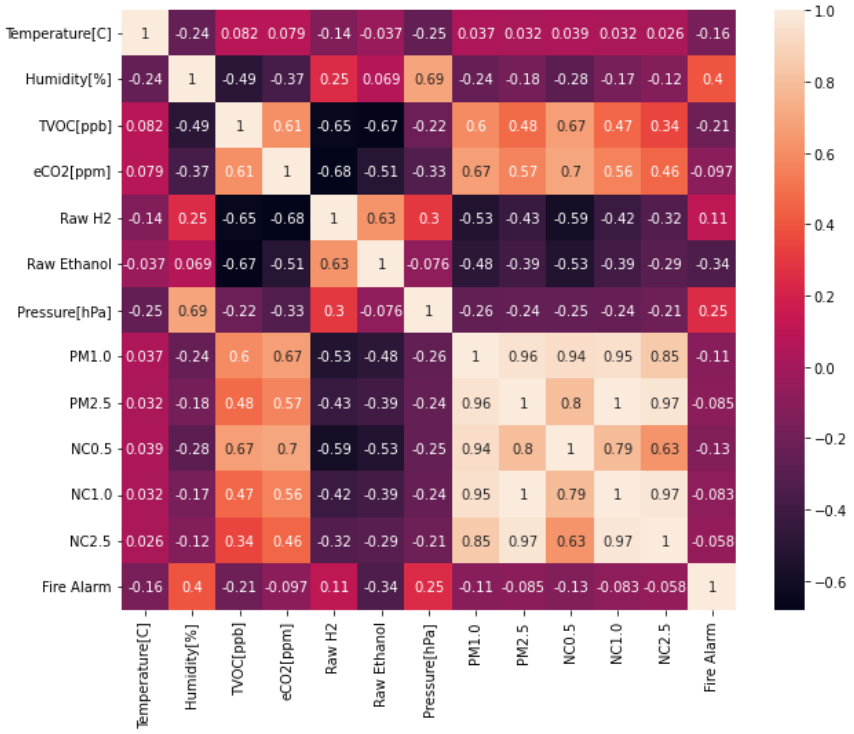
Appendix A



Appendix B



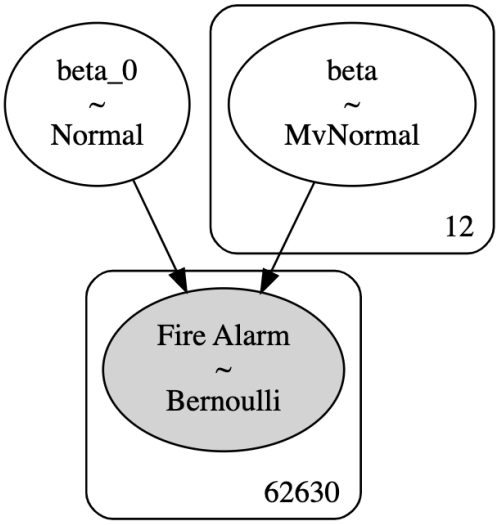
Appendix C



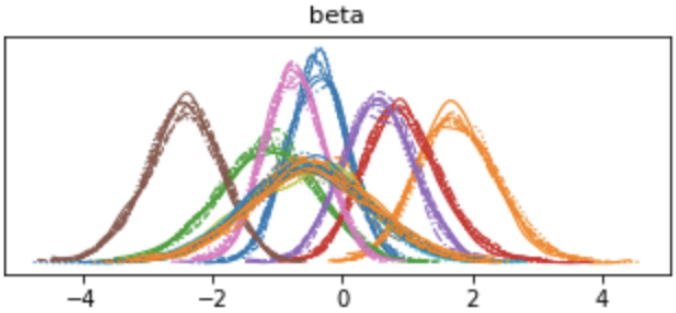
Appendix D

	Variables	Post Mean Fire	Post Mean No Fire
0	Temperature[C]	-0.103572	0.259353
1	Humidity[%]	0.252669	-0.632704
2	TVOC[ppb]	-0.135700	0.339803
3	eCO2[ppm]	-0.061300	0.153499
4	Raw H2	0.067619	-0.169324
5	Raw Ethanol	-0.215263	0.539037
6	Pressure[hPa]	0.157850	-0.395271
7	PM1.0	-0.069859	0.174934
8	PM2.5	-0.053660	0.134369
9	NC0.5	-0.080960	0.202730
10	NC1.0	-0.052340	0.131064
11	NC2.5	-0.036466	0.091314

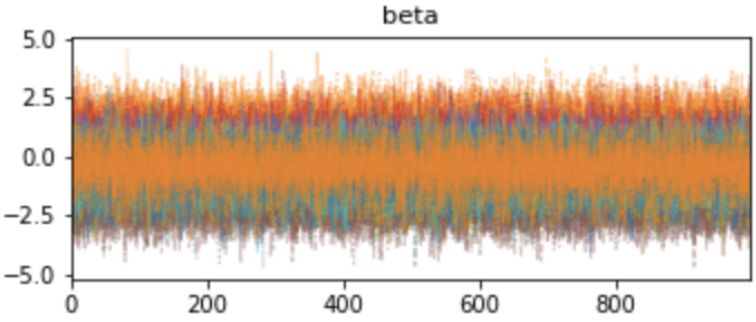
Appendix E



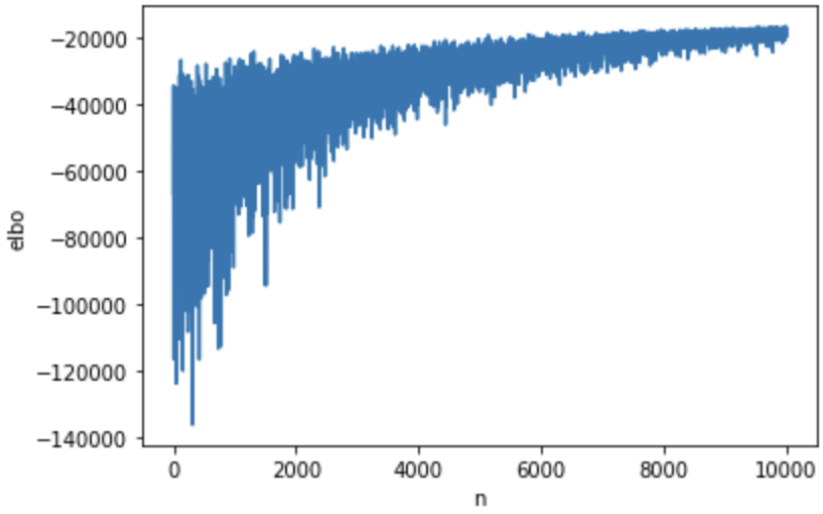
Appendix F



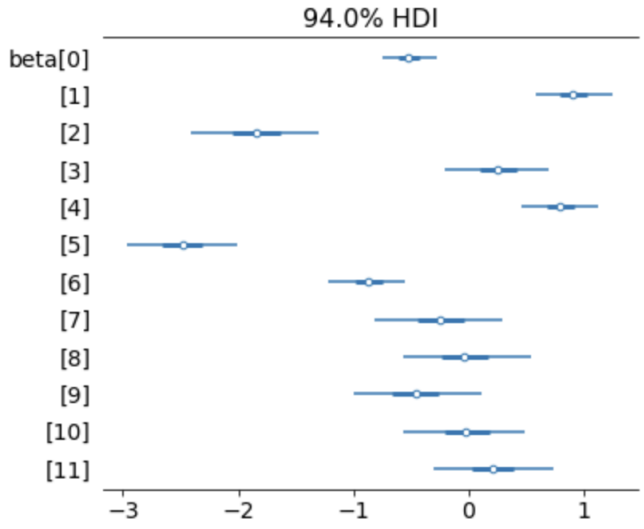
Appendix G



Appendix H



Appendix I



Appendix J

Dataset features:

- UTC*: Timestamp UTC seconds
- Temperature
- Humidity
- TVOC: Total Volatile Organic Compounds
- eCO2: co2 equivalent concentration
- Raw H2: raw molecular hydrogen
- Raw Ethanol
- Air Pressure
- PM 1.0 and PM 2.5: particulate matter size < 1.0 µm (PM1.0). 1.0 µm < 2.5 µm (PM2.5)
- CNT*: Sample counter

- NC0.5/NC1.0 and NC2.5: concentration of particulate matter. Gives the actual number of particles in the air. Also classified by the particle size: < 0.5 µm (NC0.5); 0.5 µm < 1.0 µm (NC1.0); 1.0 µm < 2.5 µm (NC2.5)
- Fire Alarm: truth value; 1 (fire present) or 0 (fire not present)

* dropped features

Works Cited

"8 Reasons Why Your Fire Alarm Goes Off Randomly." *ADT*, https://www.adt.com/resources/why-your-fire-alarm-goes-off-randomly.

"Smoke Alarm Research." *National Institute of Standards and Technology*, https://www.nist.gov/el/smoke-alarm-research.