University of Alberta

Department of mathematics and statistics

Math 509 Final Report

Case Study of The Forest Fire Data

Yuling Feng

Zhixian Yang

Zijun Zhou

December 9, 2021

Contents

1 Introduction

1.1 Background

According to the National Wildland Fire Situation Report, there are 4,182,542 hectares of forest destroyed in forest fires in 2021[Canada]. The occurrence of forest fires brings severe losses to mankind, air pollution, death of plants and animals, and even threats to human life and safety. However, traditional manual supervision and report is expensive and inefficient, so it would be beneficial to develop an efficient and low-cost detection method. Using the meteorological data collected by the automatic weather station, we wish to carry out a scientific statistical analysis of the fire situation.

Our case study focuses on modeling the probability of fire occurring in an area of the forest through the analysis of four measurements, temperature, relative humidity, rain and wind, which are very approachable through weather stations. By detecting areas with high risk, paying more attention to those specific areas during the fire season is an effective way to prevent occurences of forest fires.

The report is organized as follows. After addressing the problem in section 1, we will introduce data processing and variables used in the models in section 2. Section 3 will include two statistical models used in this study , interpretations and further analysis will be discussed in section 4. There will be a critique discussion and reference list in the end.

1.2 Problem description

We convert the fire area in the dataset into categorical data as a response variable, then take temperature, relative humidity(RH), wind speed and Rain as explanatory variables. We plan to explore whether there is a linear relationship between the fire occurring and the natural index, and build a model to predict whether a fire will occur if we know the weather conditions. We also fit a random forest model to this dataset, and compare the performance of these two models.

2 Formulation of the mathematical

2.1 The forest fire data

The dataset we used is called Forest Fires Data Set and published in 2007[Cortez and Morais, 2007]

2.1.1 FWI system

The Canadian forest fire weather index (FWI) system includes six components, Fine Fuel Moisture Code (FFMC), Duff Moisture Code (DMC), Drought Code (DC), Initial Spread Index (ISP), Buildup Index (BUI) and Fire Weather Index, that account for the effects of fuel moisture and weather conditions on fire behavior[Canada]. And those components are calculated based on the consecutive daily observations of temperature, relative humidity(RH), wind speed and 24-hour precipitation.

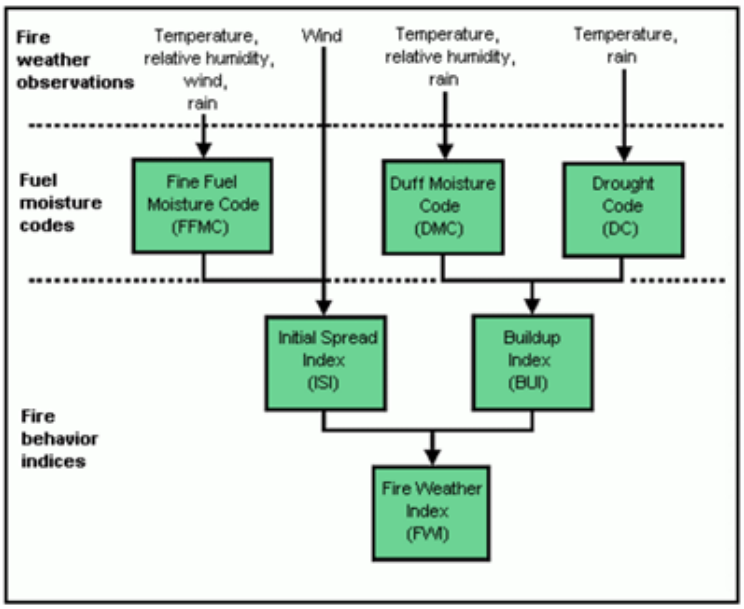


Figure 1: FWI system

2.1.2 data description and analysis

Our model will consider forest data from the Montesinho natural Park from January 2000 to December 2003. A total of 517 data were collected in the dataset. The table.1 shows all the variables in the data, and the description of the data variables as following:

The response variable

Area:

It describes the range of the area where forest fires occurred. In the dataset, there are 267 records and the remaining 250 are zero, which means that there were 267 fires.

In this case study, we redefine the value of the place where the fire occurred to be 1 and the value of the place where the fire did not occur to be 0. That is, we convert the variable "area" to

"whether there is a fire or not". From the graph we can see that the number of fires and non-fires is very close.

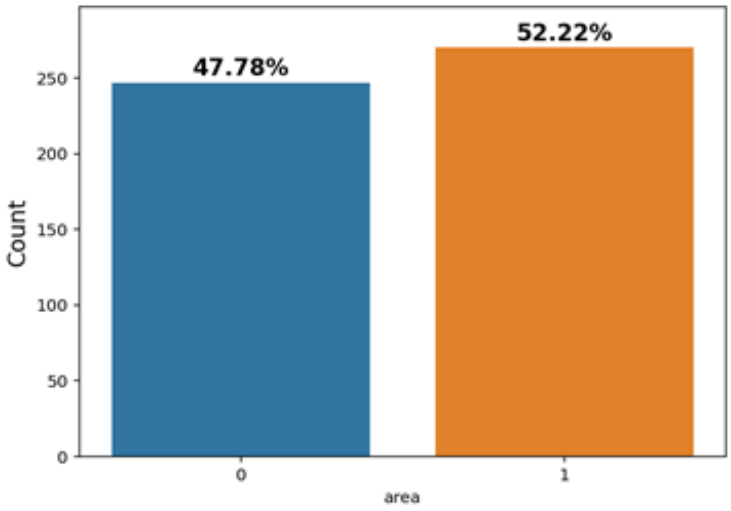


Figure 2: Proportion of area

The manipulated variables.

Those four meteorological observations were collected by the weather stations. We plot the data of independent variables into two cases 1, from all the datasets to see the distribution of four variables; 2, observe only the distribution of four variables in the area where the fire occurred. Then compare the two situations to see that there are more obvious trend differences in those variables when the fire occurred. It is clear that wind speeds are higher in the area where the fire occurred.

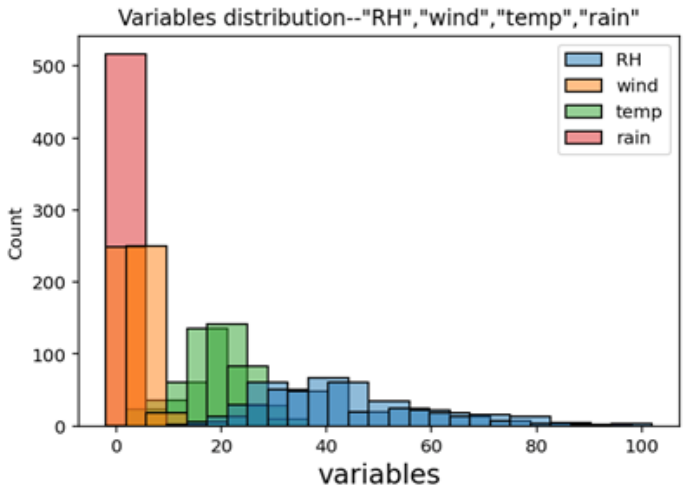


Figure 3: Variable distribution 1

RH, Wind, Temp, Rain:

The blue bar chart represents relative humidity, the green bar part represents factor temperature. They are basically following the Normal distribution. And there is not too much difference

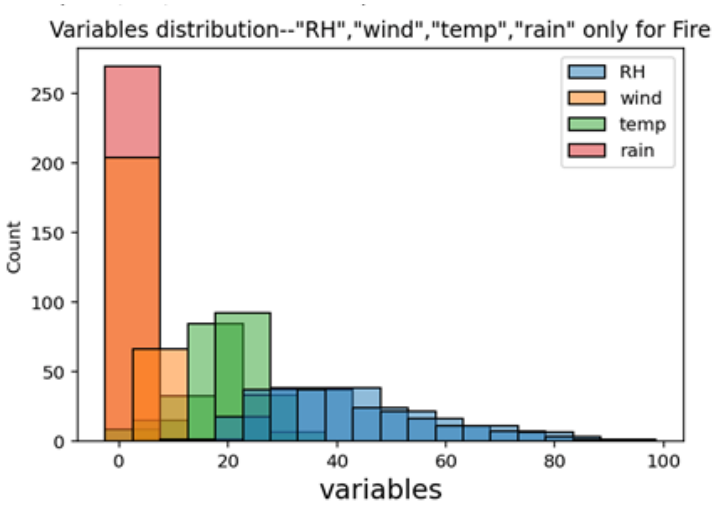


Figure 4: Variable distribution 2

between the two graphs. The pink part presenting rain indicates that the rain data are basically all concentrated in the area of 0. The orange bar chart represents wind.

In the process of modeling, we chose four most basic variables in the FWI systems, where other indexes could be calculated based on those four. It can be observed that four explanatory variables take values on very different ranges. For example, relative humidity takes values from 15 to 100 while rain falls in the range of 0 to 6.4. Therefore, we normalize the four variables so that they are all in the range of 0 to 1.

3 Solution of the problem

3.1 Introduction on modeling methods

There are mainly two models used in this project, a parametric one and a non-parametric one. After response variables are encoded to 0 and 1, classification models are needed.The performance of the linear model is not as good as expected, as a result of noises from the data. Thus we applied another non-parametric model considering that linear relationships between response variables and predictors are not significant enough.

3.2 Bayesian linear classification

Our response variables are categorical data consisting of two categories (1=fire occurred in this area, 0 = no fires). Therefore, our response variable is a Bernoulli random variable, with probability modeling by Beta distribution as prior. The formula is shown as below

$$
\begin{aligned}
\beta &= [\beta_{RH}, \beta_{temp}, \beta_{rain}, \beta_{wind}] \\
\beta_0 &\sim \text{Beta}(2, 2) \\
\beta_{RH} &\sim \text{Normal}(0.4, 0.2) \\
\beta_{wind} &\sim \text{Normal}(-0.5, 0.2) \\
\beta_{temp} &\sim \text{Normal}(0.3, 0.2) \\
\beta_{rain} &\sim \text{Normal}(-0.2, 0.2) \\
\sigma &\sim \text{Exponential}(1)
\end{aligned}
$$

We choose the parameter value for Beta distribution based on the thinking that there is always some probability of getting fired in a forest, the probability is very unlikely to be 0 or 1. Thus $\alpha = 2$ and $\beta = 2$ were chosen, according to the density graph shown below, which the density is thought to have a reasonable shape.

The parameters for the coefficients of predictors are guessed based on some basic science knowledge and a pilot research (Richards). Wind could increase the spread of fires and thus a higher value of wind would result in a higher probability of getting fire. Relative humidity and temperature usually have negative correlation since water evaporates as temperature increases. An area with high humidity is less likely to get fires thus relative humidity has a negative impact on probability of fires, and in contrast a positive value is assigned to the coefficient of temperature. The coefficient of rain has also been assigned a negative value based on the same reason.

The posterior model brings out a result similar to the prior guess. However, the posterior predictive does not have a very high accuracy, which encourages us to take a step to a non-parametric method.

3.3 Random forest

Random forest is a widely used aggregating boosting algorithm, which could also serve as a classification model. We first split the complete data set, 75% goes to the training set and the rest 25%

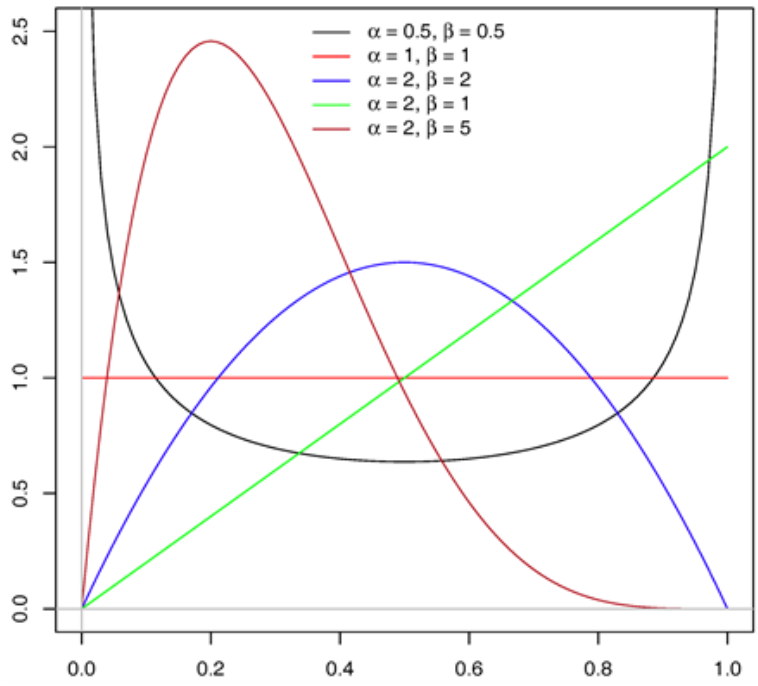


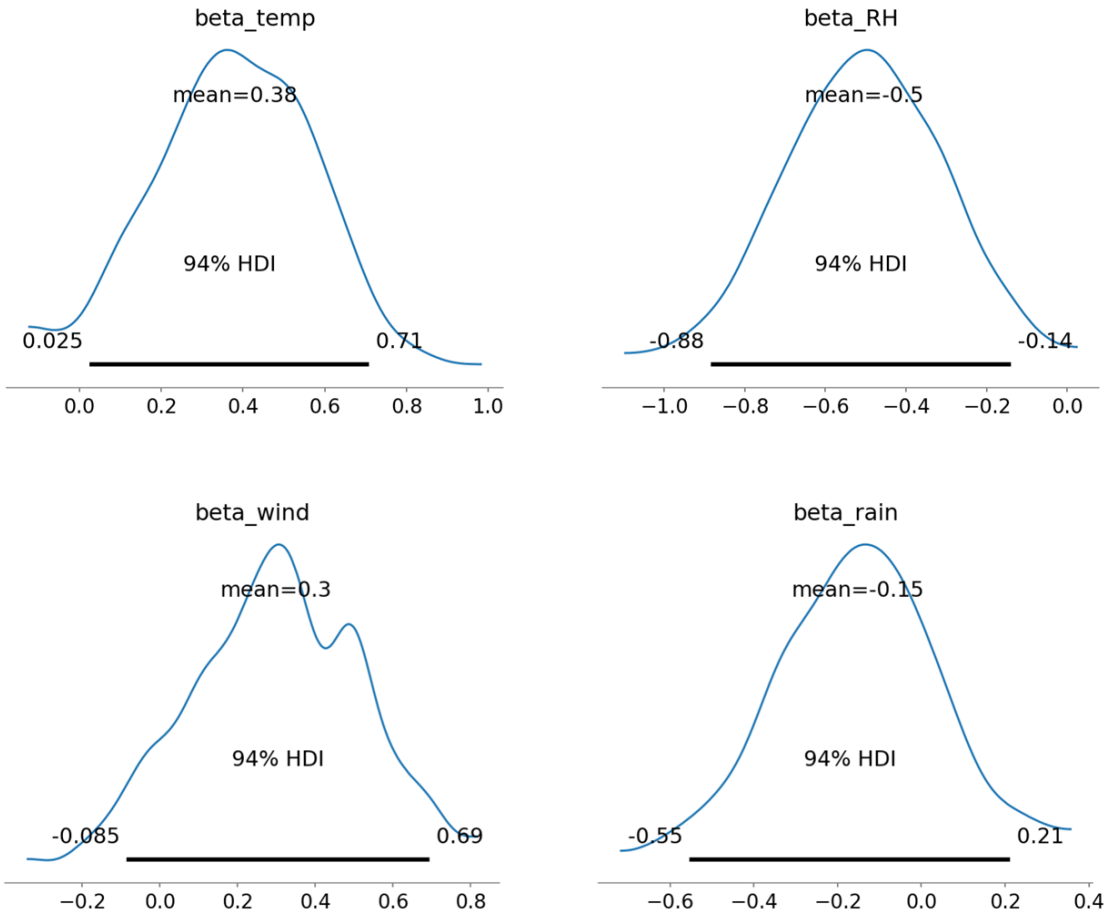
Figure 5: $\alpha = 2, \beta = 2$

	mean	sd	hdi_3%	hdi_97%	...	mcse_sd	ess_bulk	ess_tail	r_hat
beta0	0.411	0.158	0.154	0.717	...	0.012	149.0	79.0	1.04
beta_temp	0.380	0.193	0.025	0.709	...	0.009	225.0	267.0	1.02
beta_RH	-0.505	0.201	-0.884	-0.140	...	0.010	219.0	194.0	1.03
beta_wind	0.298	0.214	-0.085	0.694	...	0.010	278.0	196.0	1.00
beta_rain	-0.152	0.196	-0.553	0.212	...	0.009	231.0	170.0	1.01

Figure 6: Model coefficient

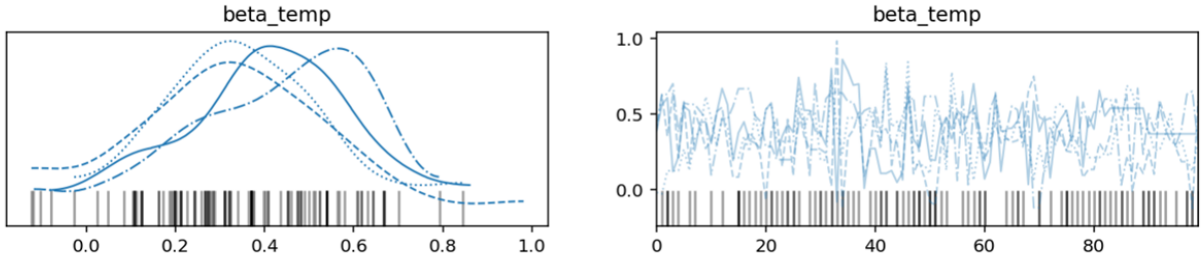
goes to the test set. The random forest model provides an accuracy of around 55%, which is still not a very high score.

4 Interpretation

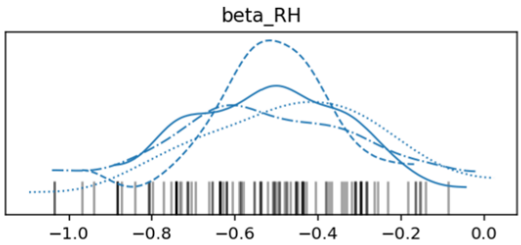


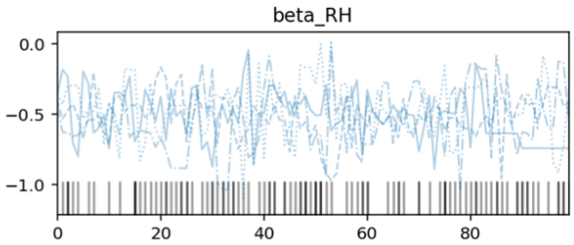
From the multiple regression model we obtained by 'pymc3', we have our estimated parameters in table 1 and the distributions of the posterior has been shown in figure 1. As can be seen from figure 1, the distribution of each variable follows normal distribution with appropriate mean and variance.

Also, the trace of posterior distribution has been shown in figure 2. We can see that 4 chains are nearly in the same shape and the black bar in the lower bound of the plot shows if the forest fire happens. As the chain trends to the mean. The frequency of accidents grows.



And our results show that β_{RH} and β_{rain} have negative numbers whether β_{temp} and β_{wind} are positive, which means that as temperature or wind increases, the probability of forest fire grows. Similarly, as relative humidity increases or rainfall rises, the possibility of forest fires decreases.





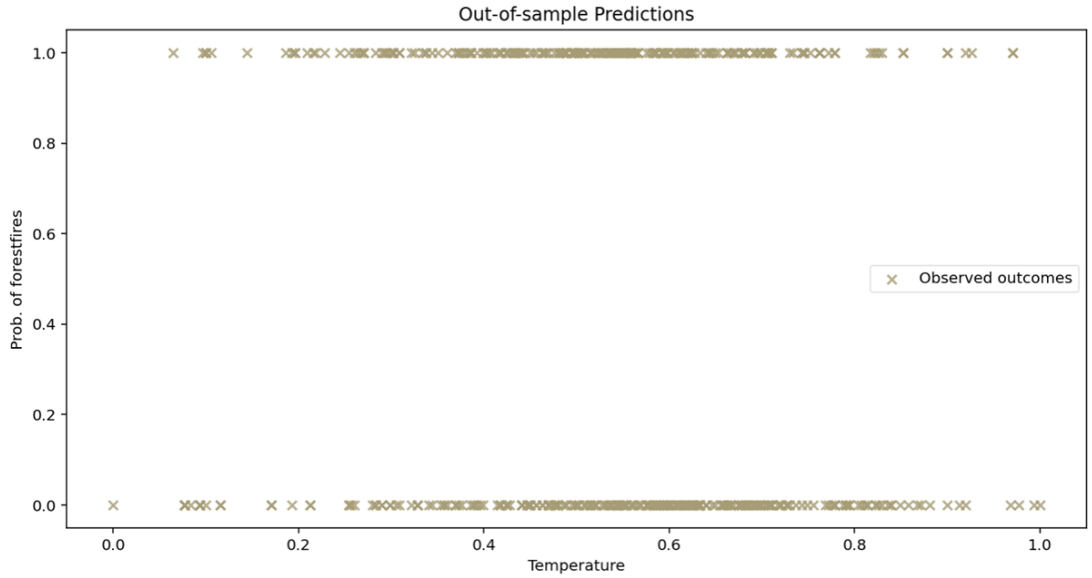
In addition, since β_{RH} has the largest absolute value(0.505), we can also conclude that the relative humidity could influence the probability of forest fires more compared to other variables and then wind makes the smallest impact on the probability based on our model.

5 Critique of the model

However, the result of the linear regression showed that the linear models did not perform well in this case study of forest fire data, although they show a reasonable direction of influence. For the multiple regression model, the accuracy of our prediction posterior is only around 51 percent, which means our model predicts does not return a convincing result.

At the same time, we also used the nonlinear model of Random Forests to analyze the data, but the results are not satisfactory. We can basically conclude that the weather conditions alone do not predict the occurrence of fires very well. We also need more information on fire prediction and control, or rely on more effective observation methods and scientific and technological measures (Liu,2016).

One of the reasons may cause the result is the data does not show that the forest fire is highly correlated with those variables. For example, Figure 2 has shown if there is any forest fire as temperature increases.



We can see that when temperature rises, then the frequency of forest fires does not increase.

In reality, it is reasonable to think that there are other relevant factors such as the time it takes for the fire to be observed and reported to fire stations, and the distance from the nearest fire stations to the ignition points. Including those variables in the model could possibly include the accuracy and prediction power of models.

References

[1] Cortez, Paulo & Morais, A.. (2007). A Data Mining Approach to Predict Forest Fires using Meteorological Data.

[2] Canada, N. R. (n.d.). Canadian wildland fire information system: Canadian forest fire weather index (FWI) system. Canadian Wildland Fire Information System — Canadian Forest Fire Weather Index (FWI) System. Retrieved December 9, 2021. (available at: https://cwfis.cfs.nrcan.gc.ca/background/summary/fwi.)

[3] Canada, N. R. (n.d.). National wildland fire situation report. Canadian Wildland Fire Information System. Retrieved December 9, 2021. (available at: https://cwfis.cfs.nrcan.gc.ca/report.)

[4] Forest fires dataset. (n.d.). Retrieved December 9, 2021. (available at: http://www3.dsi.uminho.pt/pcortez/forestfires/)

[5] Liu, Dan. "Prediction and analysis of forest fire based on machine learning." Statistic and Application 5.2 (2016): 163-171.

[6] Richards, G. D. (1988). Numerical simulation of forest fires. International journal for numerical methods in engineering, 25(2), 625-633.