2018

Study of the Relationship between a wildfire size and its cause.

The present paper tries to determine if a wildfire cause have an effect on its size. A decision tree , a decision modeling technique, is constructed to help analyze two datasets containing wildfire data statistics. The results indicates that only two causes have a significant impact on wildfires size.

Likane Anne Druide

4/15/2018

Introduction

Wildfires affect a lot of communities in the USA. Depending on a certain number of factors, the consequences can be more or less disastrous for people, the environment and properties. According to National Geographic, “On average, more than 100,000 wildfires, also called wildland fires or forest fires, clear 4 million to 5 million acres (1.6 million to 2 million hectares) of land in the U.S. every year”. The state of California is often impacted by spectacular wildfires which make the headlines. This study intends to analyze California wildfires statistics in order to try to reveal the relationship between a wildfire’s size and its cause. A decision tree modeling using R statistical software and Rattle package was built to process and analyze the datasets.

Research question

The choice of the topic of this paper came after my literature search indicated that most papers on wildfires didn’t treat the subject from the perspective I am interested in which is how wildfires causes affect their sizes.

Data sources

This paper is based on California Fire Protection Department data accessible online. Among all the available datasets, only two were selected as they seemed to be the most appropriate for the study:

1) Wildfire cause by county

This dataset gives an idea of the most frequent causes of wildfire in each country in California. Twelve (12) causes (arson, campfire, debris burning, vehicle, undetermined, electrical power, smoking, equipment use, Playing-with-fire, lightning, railroad and miscellaneous) have been identified and the dataset lists the number of fires according to their origin.

2) Wildfire size by county

Each fire that occurs is classified according to its size. Seven (7) size categories are defined in the dataset:

1) 0-0.25 acres 2) 0.26-99 acres 3) 10-99 acres 4) 100-299 acres 5) 300-999 acres 6) 1000-4999 acres and 7) 5000+ acres

Data processing

- The first data preparation step consisted in merging both datasets based on one common variable (county).
- Fire causes were categorized (numerical label assigned to each cause) so that they can be used as a predictor in the model.
- The majority of fires across all counties have a size less or equal to 0.25 acres, therefore the study only focus on this size category.
- Records that have no fire of size less or equal to 0.25 acres have been removed from the initial dataset.

Data modeling: Decision tree

Rattle was used to build a bottom up decision tree. The choice of the bottom up model seems to be the most appropriate as I am trying to use the raw data to reveal a trend in the dataset.
The target value is the fires whose size is less or to equal to 0.25 acres.

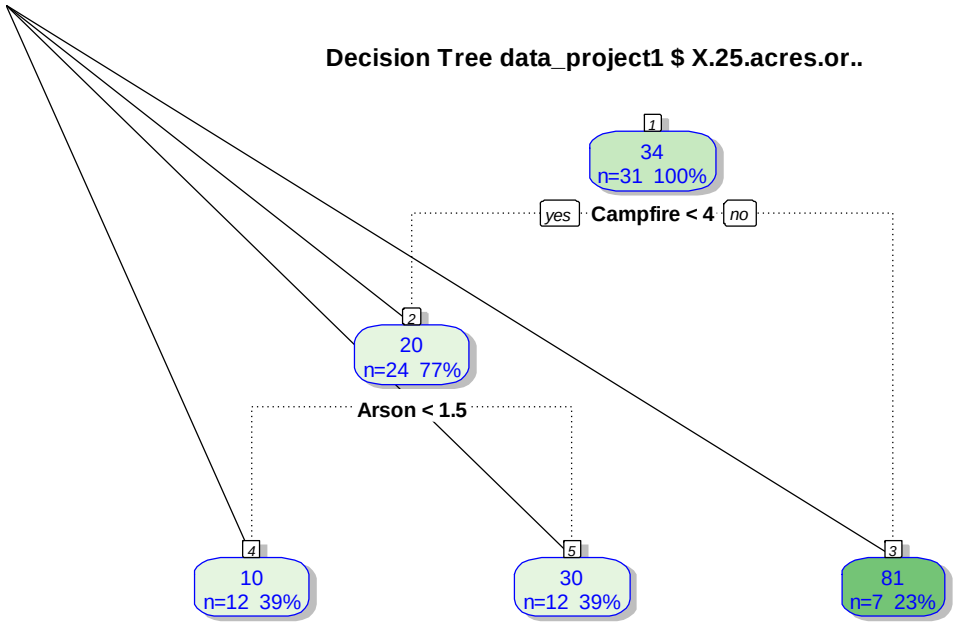
Because the target variable (size) is a numerical value, the type of decision tree is regression.

Regression tree function:

```
rpart(formula = X1 ~ ., data = crs$dataset[crs$train, c(crs$input,
    crs$target)], method = "anova", parms = list(split = "information"),
    control = rpart.control(usesurrogate = 0, maxsurrogate = 0))
```

As the function used to build the tree indicates, the model aim is to predict the X1(the number of fires with a size of 0.25 acres or less) based on the remainder of variables (fire causes) in the dataset (I decided to ignore the variable county as the final model was considering the county as the sole predictor).

Results



The tree structure is as following:

n= 31

node), split, n, deviance, yval

1) root 31 26730.970 34.03226

  2) Campfire< 4 24 6117.333 20.33333

    4) Arson< 1.5 12 1793.000 10.50000 *

    5) Arson>=1.5 12 2003.667 30.16667 *

  3) Campfire>=4 7   668.000 81.00000 *

The complexity table that reflects the model performance is below:

Revised model :

Root node error: 26731/31 = 862.29

	CP	nsplit	rel error	xerror	xstd
1	0.746162	0	1.00000	1.10767	0.21237
2	0.086816	1	0.25384	0.51752	0.16172
3	0.010000	2	0.16702	0.44399	0.16130

Initial tree:

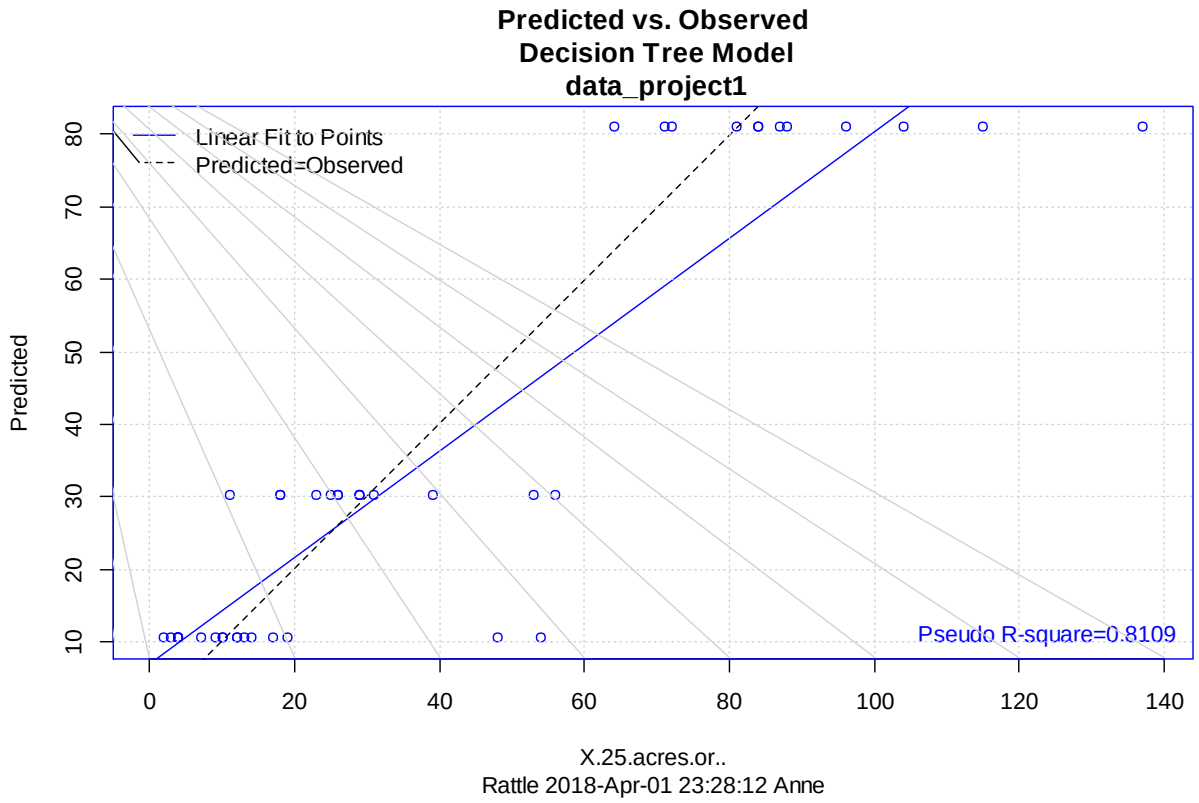
Root node error: 43761/35 = 1250.3

n= 35

	CP	nsplit	rel error	xerror	xstd
1	0.787875	0	1.00000	1.04445	0.28293
2	0.056391	1	0.21212	0.36678	0.12194
3	0.010000	2	0.15573	0.37369	0.12400

Model Evaluation

I used the Predicted vs. Observed chart option available in Rattle to evaluate the model. Pseudo-R squared which is a measure of the correlation between the predicted and observed values, is equal to 0.81. A Pseudo-R squared close to 1 indicates that the model has a good predictive characteristic and that the model fits the data.



Interpretation

"Arson" and "Campfire" are the two predictors selected by the model that have an impact on the fires whose size is less or equal to 0.25 acres.

The cross validation error rate (xerror) is reduced as we further split the tree and its lowest value is associated with the tree that best fits the data: This is tree number 3. Thus, when the number of

fires due to a campfire is greater than 4, we have the highest number (81) of fires whose size is less or equal to 0.25 acres.

Limitations:

Because Rattle only allows one target variable, it was not possible to correctly study the relationship between fires causes and their sizes. To properly answer the research question, it would have been more appropriate to use the fire causes as the model input variables to predict the number of fire in each of the 7 fire size categories, which means having to define more than one target variables.

Conclusion

This model tends to suggest that small size wildfire are affected by two causes variables namely arson and campfire. However, the greatest number of small size wildfire is associated with the campfire predictor. The information gained from the model can potentially be used to improve firefighting detection management. Depending on the extent of the damage (in terms of number of acres burned, properties damaged) caused by small fires, decision makers can determine how much money needs to be allocated to monitor areas prone to campfire.

References

California, S. O. (n.d.). Historical Wildfire Activity Statistics (Redbooks). Retrieved March 15, 2018, from http://www.fire.ca.gov/fire_protection/fire_protection_fire_info_redbooks

Learn More About Wildfires. (n.d.). Retrieved April 15, 2018, from https://www.nationalgeographic.com/environment/natural-disasters/wildfires/

Williams, G. J. (2013). Data mining with Rattle and R: The art of excavating data for knowledge discovery. New York: Springer.