Predicting ES&H Incidents

SAND2014-15949D

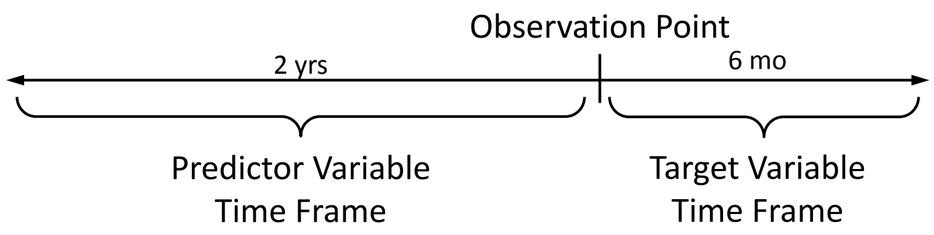
Laurel Orr, PhD Program Computer Science, University of Washington
Judy Spomer, Edward Jimenez, 9515, Software Systems R&D
Sandia National Laboratories/NM, US Department of Energy

INTRODUCTION

ES&H incidents at Sandia are of serious concern but are hard to prevent. If Sandia can identify the indicators of an organization having a safety incident in the near future, then Sandia can develop effective mitigations of those traits. By understanding the underlying factors of safety incidents occurring, we can develop a model to predict an increased risk for an organization to have safety incidents. The ultimate goal is for Sandia to use this model to take more impactful preventative measures to ensure the workforce is safe and injury free. Although this work is still in progress, we have discovered factors which have a strong relationship with the occurrence of safety incidents in the near future and will likely be used in the final model.

BUILDING A PREDICTIVE MODEL

known input variables about an Organization *(predictors)* → Final Model → probability Organization will have an incident *(target)*



Define Variables
- 4 observation points 3 months apart
- Target variable = # incidents in target time frame

Gather Data
- Brainstorm potential predictor variables
- Retrieve data from ASK (Analytics for Sandia Knowledge)
- Pull in other resources such as ES&H and HR data

Clean Data
- Does the data make sense?
- What do NULL/missing values mean?
- How do we handle outliers?

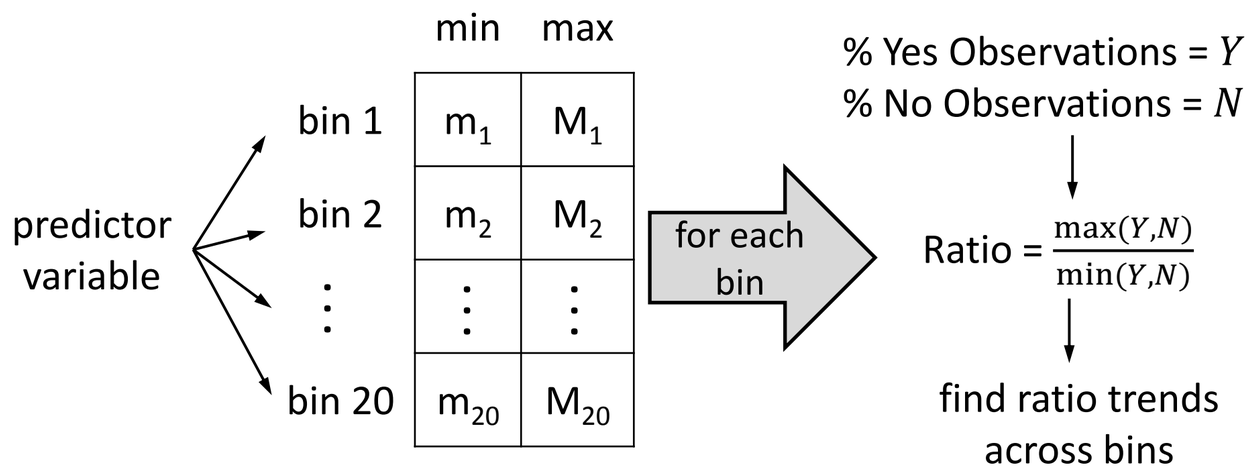
Choose Predictors
- Input variables cannot be correlated to each other
- Optionally create dummy variables
- Use bivariate analysis to find variables with a strong relationship to the target variable (see next section)

Build Models
- Partition data into test and training set
- Optionally balance training set
- Use modelling algorithms where the predictor variables act transparently → linear regression, logistic regression, decision trees

Choose Model
- Build model on training set, validate model on test set
- Choose robust model as measured by the Kolmogorov-Smirnov and Gini statistical tests

BIVARIATE ANALYSIS

Bivariate analysis examines how strong the relationship is between each predictor variable and the target variable.



PRELIMINARY BIVARIATE RESULTS

# No	# Yes	% No	% Yes	Min	Max	Ratio
6684	78	16.70	14.47	0	8	1.16
7567	128	19.01	23.75	8	12	-1.25
7113	97	17.87	18.00	12	15	-1.01
6275	64	15.76	11.87	15	18	1.33
6565	86	16.49	15.96	18	23	1.03
5610	86	14.09	15.96	23	79	-1.13

Variable A: the ratio values are erratic, which indicates no relationship between variable A and the target.

# No	# Yes	% No	% Yes	Min	Max	Ratio	
11612	115	29.17	21.34	0	2	1.37	D1
9056	108	22.75	20.04	2	5	1.14	
8408	104	21.12	19.29	5	8	1.09	D2
6648	76	16.70	14.10	8	11	1.18	
3561	114	8.94	21.15	11	115	-2.36	D3
529	22	1.33	4.08	115	375	-3.07	D4

Variable B: the ratio values trend from positive to negative, which indicates a relationship between variable B and the target.

Create Binary Dummy Variables

If $0 \le \text{var B} \le 2$ Then D1 = 1 Else D1 = 0
If $2 < \text{var B} \le 11$ Then D2 = 1 Else D2 = 0
If $11 < \text{var B} \le 115$ Then D3 = 1 Else D3 = 0
If $115 < \text{var B} \le 375$ Then D4 = 1 Else D4 = 0

FUTURE WORK

- Test various models using the variables indicated by the bivariate analysis
- Determine the strongest uncorrelated subset of variables to use in the final model
- Analyze and report findings to ES&H leadership to help them take more impactful measures to prevent future ES&H incidents

ACKNOWLEDGEMENTS

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.