Tornado - Random Forests

Arthur Lui
Sorah Kang

Department of Statistics
Brigham Young University

April 10, 2014

Introduction

Goal

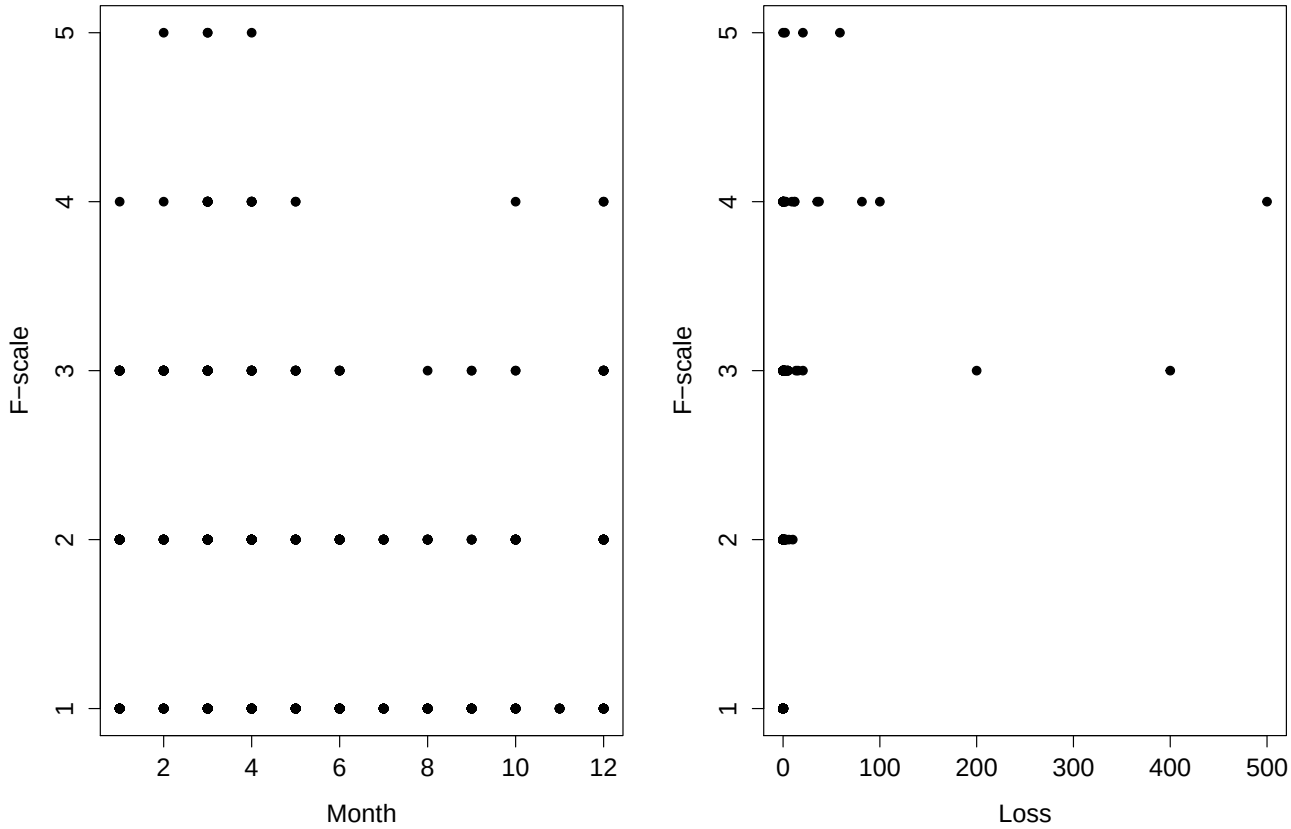
The goal of this analysis is to develop an objective model for classifying tornados

Data

Data on tornados in 2012 from the Storm Prediction Center (SPC)

- Fscale - Fujita Scale (Response - subjective)
- Number - SPC Tornado Number
- Month - Numeric Month Value
- Day - Day of the Month
- Time - Time Tornado First Reported
- Loss - Rounded Total Dollar Property Loss (in millions)
- CropLoss - Rounded Total Dollar Crop Loss (in millions)
- Length - Length of impact (in miles)
- Width - Width of impact (in yards)

Data



F-Scale	0	1	2	3	4	5
Freq	578	242	100	32	5	0

Data Cleaning

23 tornados were repeated in the data

Data Cleaning

23 tornados were repeated in the data

	Number	Date	State	F-scale	Injuries	Loss	Length
120	359692	2/29/12	IL	2	0	0.05	8.41
121	359692	2/29/12	IL	2	5	0.30	25.12
122	359692	2/29/12	KY	2	5	0.25	26.71

Model

Random Forest - The idea is to create a multitude of trees (forest)

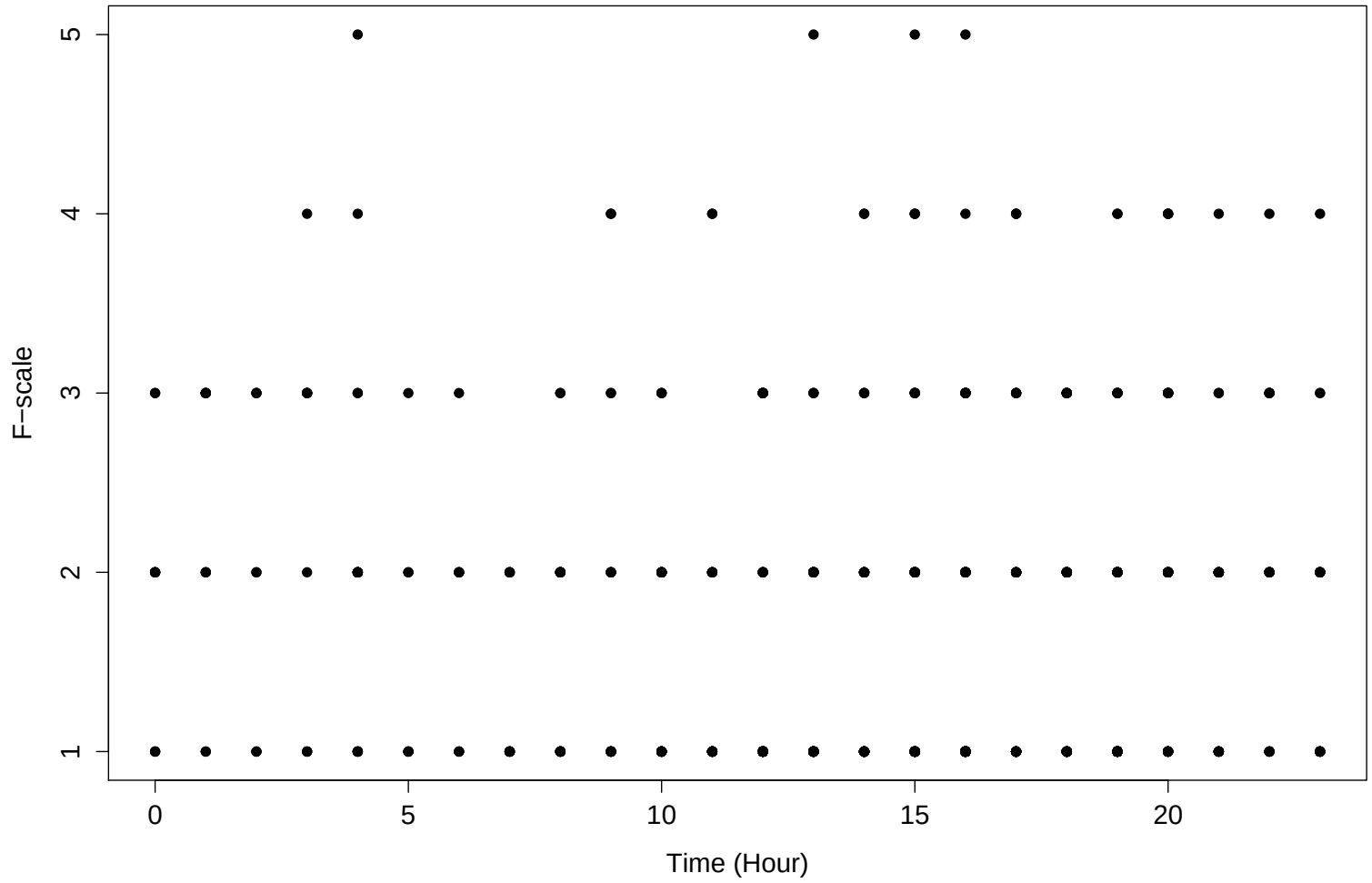
1. Take a bootstrap sample of size n.
2. At each split, randomly sample $m < P$ variables
3. Build a tree based on each set
    1. At each node, identify variable that has most correlation with output
    2. Identify a cut point that leads to the greatest reduction in error

Random Forests

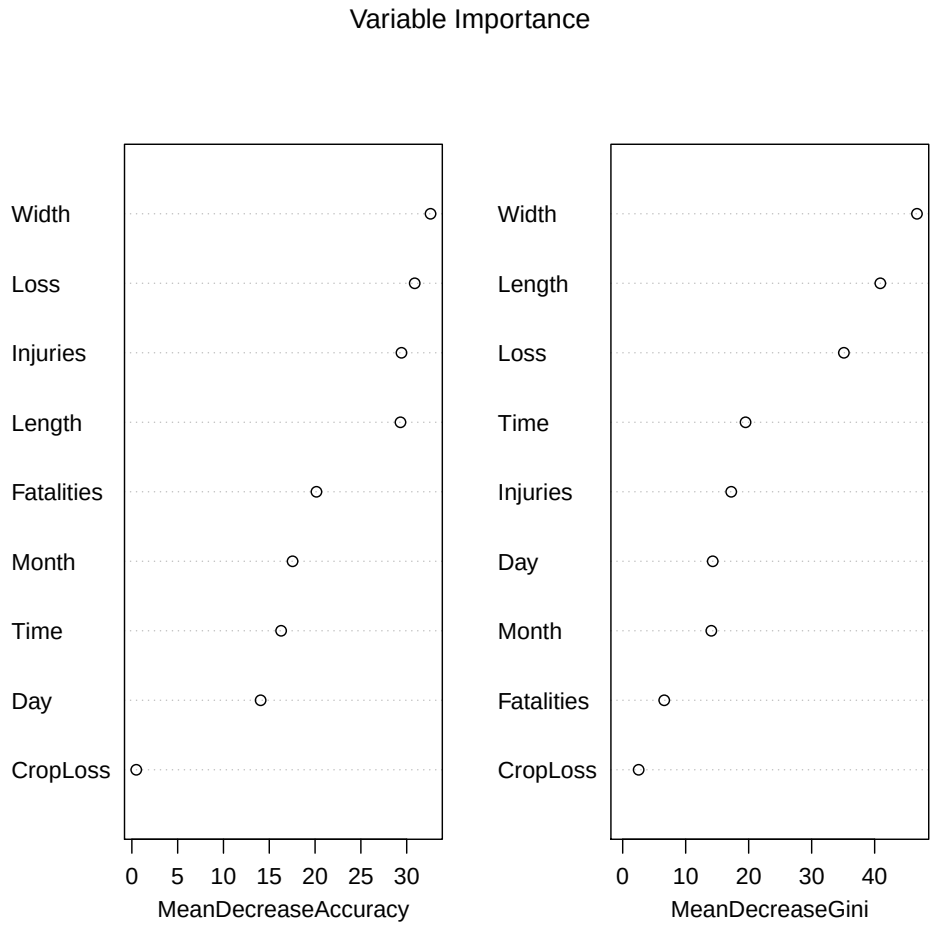
+ Better predictive power than trees
  - Choosing a subset of variables decorrelates the trees
+ Gives us an objective way of classifying tornados
- However, random forests are not very interpretable

Model Assumptions

For classification trees, we assume data is non-linear

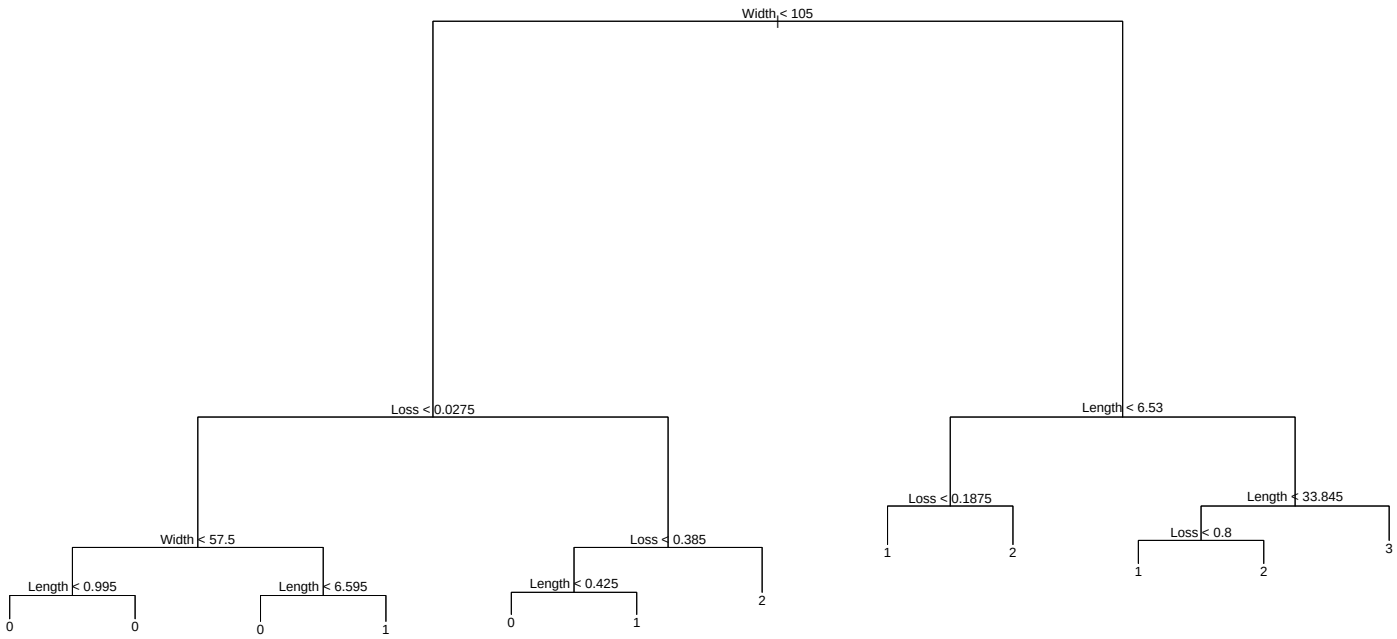


Most Important Variables



The most important variables in predicting Fscale are: Loss, Width, and Length.

Tree



How would a scientist use this model?

Confusion Matrix

	0	1	2	3	4	class.error
0	515.00	58.00	3.00	0.00	0.00	0.11
1	88.00	135.00	17.00	1.00	0.00	0.44
2	4.00	43.00	44.00	4.00	0.00	0.54
3	0.00	5.00	15.00	4.00	0.00	0.83
4	0.00	1.00	1.00	2.00	0.00	1.00

Note: Error rate is lower for small Fscales

Conclusions

- Error Rate $\approx 26\%$
- Objective and takes into account past data
- Predicts low Fscales well (because more data)

Future

- Need more data for Fscale $= 4$
- Compare different methods (e.g. trees, bagging, etc.)

Teamwork

- Did Analysis together
- Sorah: Introduction, Model, Data
- Arthur: Results, Conclusions, Future