An exploration of data: prediction, modeling and displays

Karsten Maurer

Department of Statistics
Miami University
Oxford, OH 45056
maurerkt@miamioh.edu

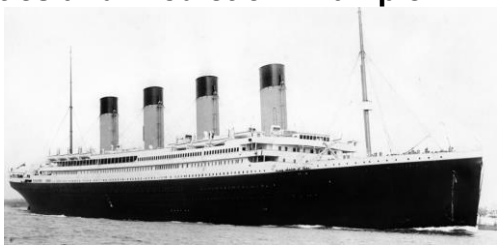
URL: kmaurer.github.io

Department Social Media

Twitter: @STADeprMiamiOH @statsandstories

Podcast: www.statsandstories.net

Statistics and Prediction Example



Background: In the early morning of April 15, 1912, the RMS Titanic sunk after colliding with an iceberg in the North Atlantic during her maiden voyage from the Samhampton, UK to New York, USA. Roughly 2/3 of the passengers and crew did not survive this accident.

Data: In the paper, The "Unusual Episode" Data Revisited published in the *Journal of Statistics Education* vol.3, no.3 (1995), records for 2201 passengers and crew were recorded with their ticket status (the Class variable), Age (categorized as Adult/Child), Gender (Female/Male) and whether they survived the sinking. 15 of the 2201 passengers/crew were randomly removed from the record and summary tables of the remaining 2186 passengers/crew is included below.

Questions:

1. Which variables appear to influence a person's survival?
2. On the next page is a list of the 15 people removed from the record. Your goal is to:
   A) Predict whether each of the 15 people survived.
   B) Assign a probability/percentage on the likelihood they survived.

Class	
1st	324
2nd	283
3rd	701
Crew	878

Age	
Adult	2078
Child	108

Gender	
Female	465
Male	1721

Survived	
No	1483
Yes	703

	Survived	
Class	No	Yes
1st	122	202
2nd	167	116
3rd	526	175
Crew	668	210

	Survived	
Age	No	Yes
Adult	1431	647
Child	52	56

	Survived	
Gender	No	Yes
Female	126	339
Male	1357	364

	Gender	
Age	Female	Male
Adult	421	1657
Child	44	64

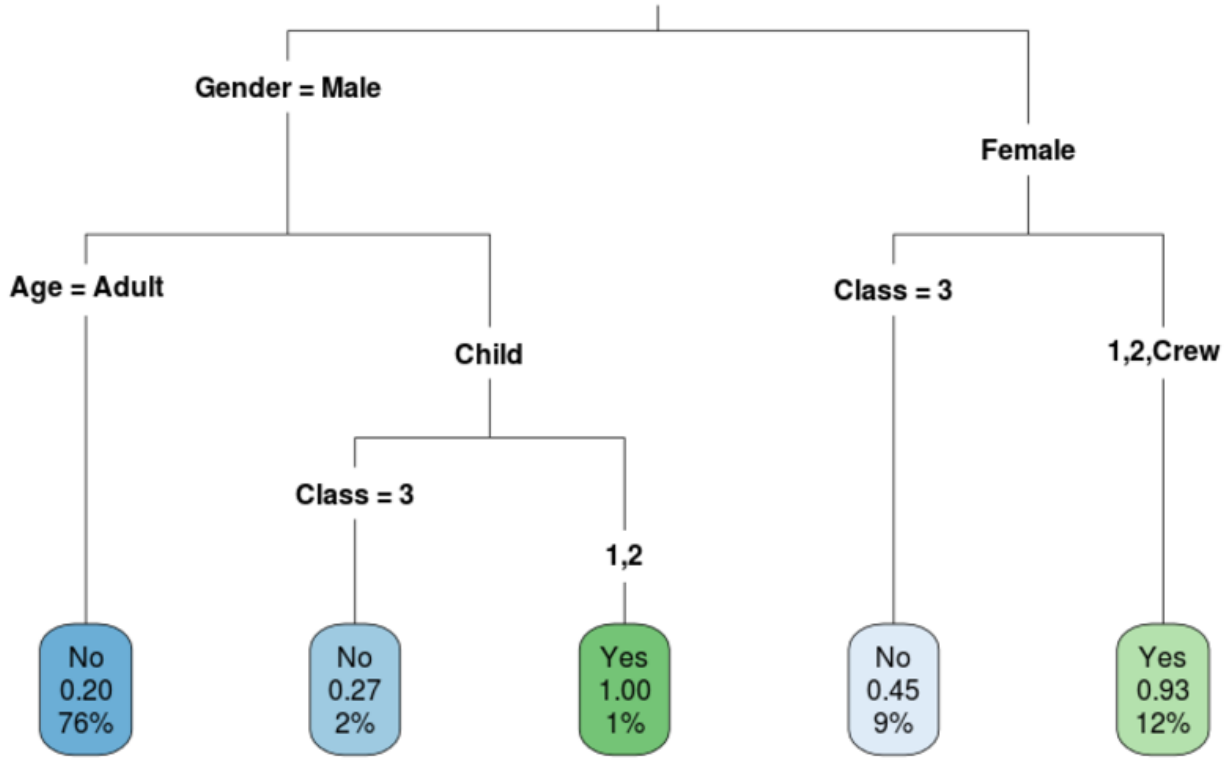
	Age	
Class	Adult	Child
1st	318	6
2nd	260	23
3rd	622	79
Crew	878	0

	Gender	
Class	Female	Male
1st	144	180
2nd	104	179
3rd	194	507
Crew	23	855

			Survived	
Class	Age	Gender	No	Yes
1st	Adult	Female	4	139
		Male	118	57
	Child	Female	0	1
		Male	0	5
2nd	Adult	Female	13	79
		Male	154	14
	Child	Female	0	12
		Male	0	11
3rd	Adult	Female	89	74
		Male	385	74
	Child	Female	17	14
		Male	35	13
Crew	Adult	Female	3	20
		Male	665	190
	Child	Female	0	0
		Male	0	0

Person	Class	Age	Gender
219	1st	Adult	Female
566	2nd	Adult	Female
602	2nd	Child	Female
633	3rd	Adult	Male
815	3rd	Adult	Male
866	3rd	Adult	Male
1104	3rd	Adult	Female
1122	3rd	Adult	Female
1402	Crew	Adult	Male
1407	Crew	Adult	Male
1672	Crew	Adult	Male
1854	Crew	Adult	Male
2025	Crew	Adult	Male
2097	Crew	Adult	Male
2135	Crew	Adult	Male

Suppose we built a statistical model … a classification tree was produced below based on a training set of 2186 passengers. (STA 333, STA 467)



This could be applied to the test set of 15 passengers that were sampled.

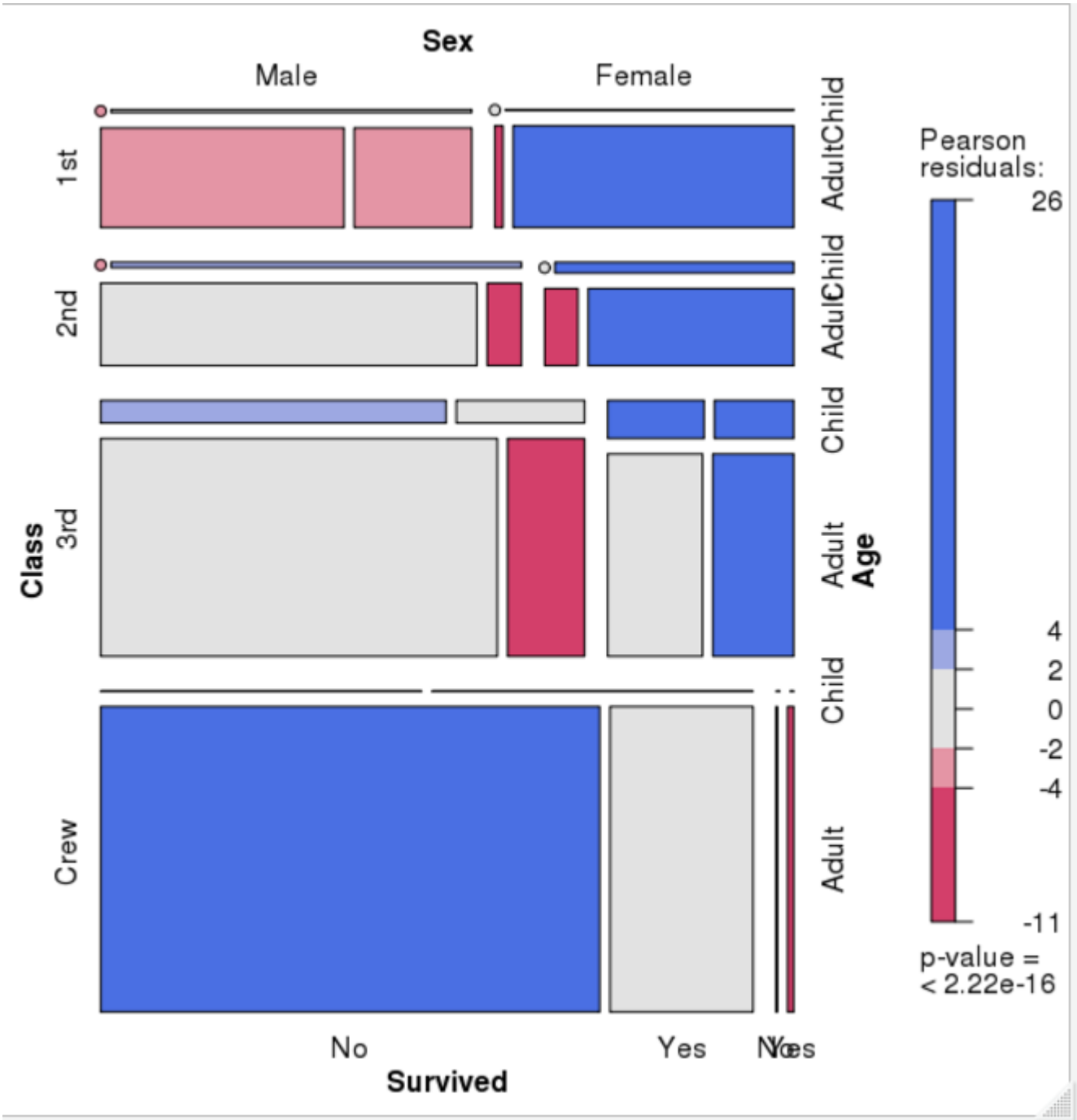
Predictions based on classification tree

Person	Class	Age	Gender	P(Not)	P(Survived)	Predict
219	1st	Adult	Female	0.07380074	0.9261993	Yes
566	2nd	Adult	Female	0.07380074	0.9261993	Yes
602	2nd	Child	Female	0.07380074	0.9261993	Yes
633	3rd	Adult	Male	0.79782740	0.2021726	No
815	3rd	Adult	Male	0.79782740	0.2021726	No
866	3rd	Adult	Male	0.79782740	0.2021726	No
1104	3rd	Adult	Female	0.54639175	0.4536082	No
1122	3rd	Adult	Female	0.54639175	0.4536082	No
1402	Crew	Adult	Male	0.79782740	0.2021726	No
1407	Crew	Adult	Male	0.79782740	0.2021726	No
1672	Crew	Adult	Male	0.79782740	0.2021726	No
1854	Crew	Adult	Male	0.79782740	0.2021726	No
2025	Crew	Adult	Male	0.79782740	0.2021726	No
2097	Crew	Adult	Male	0.79782740	0.2021726	No
2135	Crew	Adult	Male	0.79782740	0.2021726	No

Person	Class	Age	Gender	P(Not)	P(Survived)	Prediction	Truth
219	1st	Adult	Female	0.07380074	0.9261993	Yes	**Yes**
566	2nd	Adult	Female	0.07380074	0.9261993	Yes	**Yes**
602	2nd	Child	Female	0.07380074	0.9261993	Yes	**Yes**
633	3rd	Adult	Male	0.79782740	0.2021726	No	**Yes**
815	3rd	Adult	Male	0.79782740	0.2021726	No	**No**
866	3rd	Adult	Male	0.79782740	0.2021726	No	**No**
1104	3rd	Adult	Female	0.54639175	0.4536082	No	**Yes**
1122	3rd	Adult	Female	0.54639175	0.4536082	No	**Yes**
1402	Crew	Adult	Male	0.79782740	0.2021726	No	**Yes**
1407	Crew	Adult	Male	0.79782740	0.2021726	No	**Yes**
1672	Crew	Adult	Male	0.79782740	0.2021726	No	**No**
1854	Crew	Adult	Male	0.79782740	0.2021726	No	**No**
2025	Crew	Adult	Male	0.79782740	0.2021726	No	**No**
2097	Crew	Adult	Male	0.79782740	0.2021726	No	**No**
2135	Crew	Adult	Male	0.79782740	0.2021726	No	**No**

Our predictions of the 15 passengers that were sampled wasn't perfect (10 of 15 classified correctly).

Tables can be tough to process. Can we visualize this? (STA 404)



Visual cues in this Mosaic Plot?

- Size of boxes
- Color of boxes

Visualizing data …

The Joy of Statistics – Hans Rosling
https://www.youtube.com/watch?v=jbkSRLYSojo

As you watch this video, please record the following information:

What variables were presented?

What graphical characteristic (aesthetic trait) was mapped to each variable?

Rstudio.miamioh.edu -> exploring gapminder data
http://kmaurer.github.io/documents/Data-visualization-exploration.R

Studying at Miami University

- Math & Stat Degrees (B.S. Math & Stat, B.S. Stat)
  - Foundation in mathematics
  - Statistical modeling
  - Data handling and visualization
- Analytics Co-Major
  - Complements the B.S. Math & Statistics & B.S. Statistics very well
- Actuarial Science Minor (actuarial science club: Dr. Miljkovic)
  - Complements B.S. degrees and satisfies related hours & thematic sequences
- Miami University StatHawks (partners with Pi Mu Epsilon for some activities)
  - Student Chapter of the American Statistical Association
  - Can join on the Hub - Events throughout the fall (movie night, trivia, speakers)
- Center for Analytics and Data Science (CADS)
  - DataFest - weekend of April 6-8, 2018

THANK YOU!

Questions?