

PROFESSIONAL & CONTINUING EDUCATION

UNIVERSITY *of* WASHINGTON

Machine Learning Techniques

DATASCI 420

Lesson 05-02 Decision Trees, Part 2



Example: Tragedy of the Titanic



Women and
children first?

W

AutoSaveOff

Titanic_Dataset.xlsx

Table Tools

Kristin Tolle

File

Home

Insert

Draw

Page Layout

Formulas

Data

Review

View

Design

Tell me what you want to do

C14

Start with a Data Definition

| Feature | Data Description | Variable Type |
|----------|--|----------------------|
| survival | Survived (0 = no; 1 = yes) | Dependent variable |
| pclass | Ticket class 1 = 1st, 2 = 2nd, 3 = 3rd | Independent variable |
| name | Passenger's name as given | Relevance? |
| sex | Gender: Male or Female | Independent variable |
| age | Age | Independent variable |
| sibsp | Count of siblings and/or spouse aboard | Independent variable |
| parch | Count of parents or children aboard | Independent variable |
| ticket | Ticket number | Relevance? |
| fare | Passenger Fare | Independent variable |
| cabin | Cabin number | Relevance? |
| embarked | Port of departure (c = Cherbourg, q=Queenstown, s=Southampton) | Relevance? |
| boat | Rescue boat number | Relevance? |
| home | Home city of the passenger and ultimate (assumed) destination | Relevance? |
| ... | | |

Choose wisely: More variables in the model can negatively affect compute time and potentially accuracy

First Steps... Look for Problems and Predictors

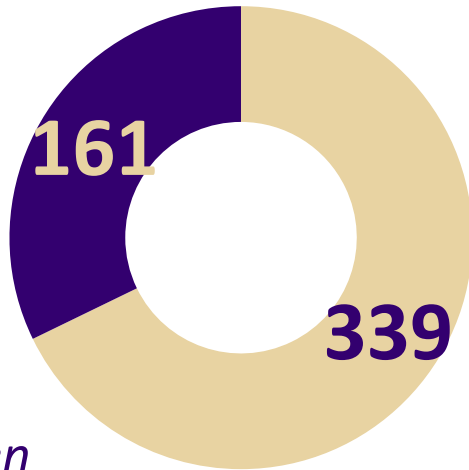
| | pclass | survived | name | sex | age | sibsp | parch |
|-----|--------|----------|--|--------|-----|-------|-------|
| 261 | 1 | 1 | Seward, Mr. Frederic Kimber | male | 34 | 0 | 0 |
| 262 | 1 | 1 | Shutes, Miss. Elizabeth W | female | 40 | 0 | 0 |
| 263 | 1 | 1 | Silverthorne, Mr. Spencer Victor | male | 35 | 0 | 0 |
| 264 | 1 | 0 | Silvey, Mr. William Baird | male | 50 | 1 | 0 |
| 265 | 1 | 1 | Silvey, Mrs. William Baird (Alice Munger) | female | 39 | 1 | 0 |
| 266 | 1 | 1 | Simonius-Blumer, Col. Oberst Alfons | male | 56 | 0 | 0 |
| 267 | 1 | 1 | Sloper, Mr. William Thompson | male | 28 | 0 | 0 |
| 268 | 1 | 0 | Smart, Mr. John Montgomery | male | 56 | 0 | 0 |
| 269 | 1 | 0 | Smith, Mr. James Clinch | male | 56 | 0 | 0 |
| 270 | 1 | 0 | Smith, Mr. Lucien Philip | male | 24 | 1 | 0 |
| 271 | 1 | 0 | Smith, Mr. Richard William | male | | 0 | 0 |
| 272 | 1 | 1 | Smith, Mrs. Lucien Philip (Mary Eloise Hughes) | female | 18 | 1 | 0 |
| 273 | 1 | 1 | Snyder, Mr. John Pillsbury | male | 24 | 1 | 0 |
| 274 | 1 | 1 | Snyder, Mrs. John Pillsbury (Nelle Stevenson) | female | 23 | 1 | 0 |
| 275 | 1 | 1 | Spedden, Master. Robert Douglas | male | 6 | 0 | 2 |
| 276 | 1 | 1 | Spedden, Mr. Frederic Oakley | male | 45 | 1 | 1 |
| 277 | 1 | 1 | Spedden, Mrs. Frederic Oakley (Margaretta Corning Stone) | female | 40 | 1 | 1 |
| 278 | 1 | 0 | Spencer, Mr. William Augustus | male | 57 | 1 | 0 |
| 279 | 1 | 1 | Spencer, Mrs. William Augustus (Marie Eugenie) | female | | 1 | 0 |
| 280 | 1 | 1 | Stahelin-Maeglin, Dr. Max | male | 32 | 0 | 0 |
| 281 | 1 | 0 | Stead, Mr. William Thomas | male | 62 | 0 | 0 |
| 282 | 1 | 1 | Stengel, Mr. Charles Emil Henry | male | 54 | 1 | 0 |
| 283 | 1 | 1 | Stengel, Mrs. Charles Emil Henry (Annie May Morris) | female | 43 | 1 | 0 |
| 284 | 1 | 1 | Stephenson, Mrs. Walter Bertram (Martha Eustis) | female | 52 | 1 | 0 |
| 285 | 1 | 0 | Stewart, Mr. Albert A | male | | 0 | 0 |
| 286 | 1 | 1 | Stone, Mrs. George Nelson (Martha Evelyn) | female | 62 | 0 | 0 |
| 287 | 1 | 0 | Straus, Mr. Isidor | male | 67 | 1 | 0 |
| 288 | 1 | 0 | Straus, Mrs. Isidor (Rosalie Ida Blun) | female | 63 | 1 | 0 |
| 289 | 1 | 0 | Sutton, Mr. Frederick | male | 61 | 0 | 0 |
| 290 | 1 | 1 | Swift, Mrs. Frederick Joel (Margaret Welles Barron) | female | 48 | 0 | 0 |
| 291 | 1 | 1 | Taussig, Miss. Ruth | female | 18 | 0 | 2 |
| 292 | 1 | 0 | Taussig, Mr. Emil | male | 52 | 1 | 1 |
| 293 | 1 | 1 | Taussig, Mrs. Emil (Tillie Mandelbaum) | female | 39 | 1 | 1 |

Look for Missing Values
e.g., Age...

Make sure you
understand column
heading

Survival by Gender (Passengers)

Survival Rate



N = 1324

12% of men

26% of women

More Women
Survived

**IF sex='female' THEN survive=yes
ELSE IF sex='male' THEN survive = no**

confusion matrix

no yes<-- classified as

| | | |
|-----|-----|-----|
| 468 | 109 | no |
| 81 | 233 | yes |

$(468 + 233) / (468 + 109 + 81 + 233) = 79\%$ correct (and 21% incorrect)

Not bad!!

Survival Rate by Gender and Class



Regardless of class,
more women survived;
However...

From this view it seems
that class mattered
more than gender

**IF pclass='1' THEN survive=yes
ELSE IF pclass='2' THEN survive=yes
ELSE IF pclass='3' THEN survive=no**

confusion matrix

no yes<-- classified as
372 119 | no
177 223 | yes

$(372 + 223) / (372 + 119 + 223 + 177) = 67\%$ correct (and 33% incorrect)

A little worse



Jupyter Notebook for the Titanic