Titanic Machine Learning Project

Jaleel Savoy, 11/2016

Explanation of My Process

- I took into consideration the social norms of the time period
- The initial assumption was that Sex, Class, and Age were the most important explanatory variables to determine the response variable of Survival
- I chose to use a recursive partitioning analysis method and create a decision/regression tree model
    - The regression tree model was chosen due to its easy digestibility for readers
    - The models were created to be simple and efficient to avoid overfitting the model
    - The models were cross-validated before submission for scoring

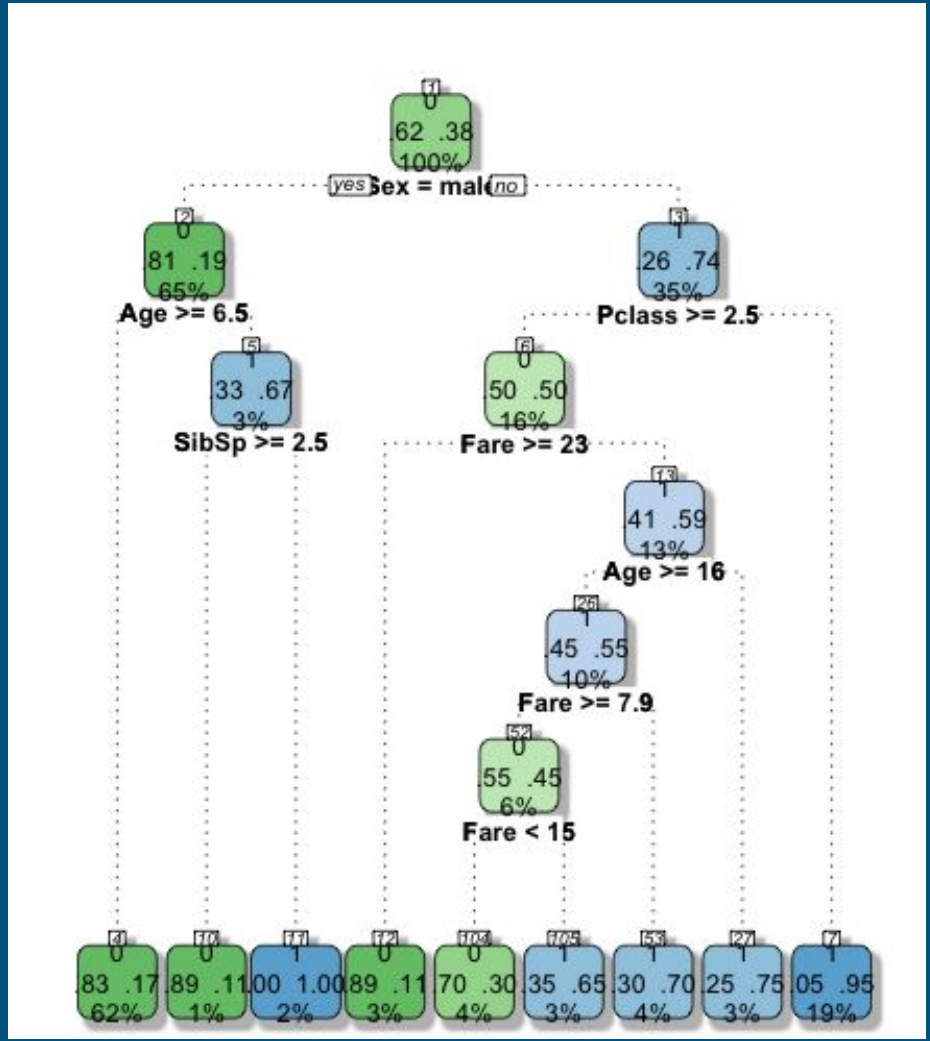
Summary of Analysis and Results

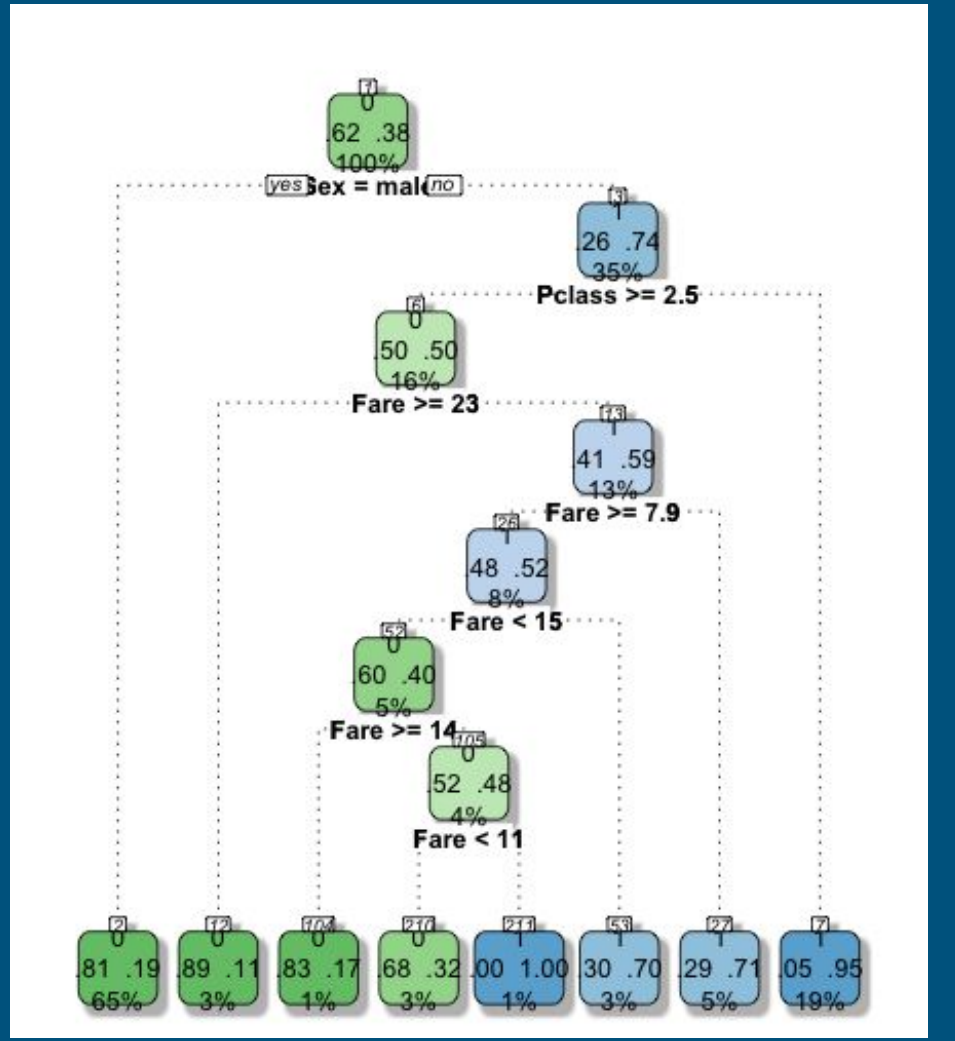
- The 1st model had a prediction accuracy of 76%
  - Explanatory variables: Sex, Age, Fare, Pclass, and SibSp (# of siblings and spouses)

- The 2nd model had a prediction accuracy of ~77.512%
  - Explanatory variables: Sex, Fare (ticket price), and Pclass (Socioeconomic class)
  - Originally also had Age as an explanatory variable, but was removed for more accuracy
    - Very insightful that removing age slightly improved the model

- Additional questions to now consider would be why passengers of certain fares survived and others did not
  - The splits for fare are very interesting

- The shortfalls or limitations of the work include:
  - Relatively quick creation of the model
  - Limited amount of cross-validation performed
  - Possibility of choosing an overly simple model


1
0
.62 .38
100%
yes
Sex = mal
no
2
0
.81 .19
65%
Age >= 6.5
3
1
.26 .74
35%
Pclass >= 2.5
5
1
.33 .67
3%
SibSp >= 2.5
6
0
.50 .50
16%
Fare >= 23
13
1
.41 .59
13%
Age >= 16
26
.45 .55
10%
Fare >= 7.9
52
0
.55 .45
6%
Fare < 15
4
0
.83 .17
62%
10
0
.89 .11
1%
11
1
.00 1.00
2%
12
0
.89 .11
3%
104
0
.70 .30
4%
105
1
.35 .65
3%
53
1
.30 .70
4%
27
1
.25 .75
3%
7
1
.05 .95
19%


1
0
.62 .38
100%
yes
Sex = mal
no
3
1
.26 .74
35%
Pclass >= 2.5
6
0
.50 .50
16%
Fare >= 23
13
1
.41 .59
13%
Fare >= 7.9
26
1
.48 .52
8%
Fare < 15
52
0
.60 .40
5%
Fare >= 14
105
0
.52 .48
4%
Fare < 11
2
0
.81 .19
65%
12
0
.89 .11
3%
104
0
.83 .17
1%
210
0
.68 .32
3%
211
1
.00 1.00
1%
53
1
.30 .70
3%
27
1
.29 .71
5%
7
1
.05 .95
19%