Titanic Survivors

James Whedbee

October 23, 2015

Overview

The following is an attempt to predict survival of passengers on the Titanic. The data available for use is as follows:

```
## 'data.frame':    891 obs. of  12 variables:
##  $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Survived   : int  0 1 1 1 0 0 0 0 1 1 ...
##  $ Pclass     : int  3 1 3 1 3 3 1 3 3 2 ...
##  $ Name       : Factor w/ 891 levels "Abbing, Mr. Anthony",..: 109 191 358 277 16 559 520 629 417 58
##  $ Sex        : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
##  $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
##  $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
##  $ Ticket     : Factor w/ 681 levels "110152","110413",..: 524 597 670 50 473 276 86 396 345 133 ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin      : Factor w/ 148 levels "","A10","A14",..: 1 83 1 57 1 1 131 1 1 1 ...
##  $ Embarked   : Factor w/ 4 levels "","C","Q","S": 4 2 4 4 4 3 4 4 4 2 ...
```

SPECIAL NOTES:

Pclass is a proxy for socio-economic status: 1st ~ Upper; 2nd ~ Middle; 3rd ~ Lower

SibSp refers to the number of siblings or spouses

Parch refers to the number of parents or children

Since age was not available for all passengers, missing age values were projected using a linear model composed of number of siblings/spouses, number of parents/children, and class. For the rest of the paper, any mention of Age is actually referring to the projected age.

Primary predictors - Sex and Class

Starting from nothing, correlations with survival will give us a rough sense of the relative importance of variables.

```
##         Sex      Pclass        Fare    Embarked       Parch         Age
## -0.53882559 -0.35965268  0.26818862 -0.18997851  0.09331701 -0.07722109
##     projAge       SibSp
## -0.07722109 -0.01735836
```

Sex, Pclass, and Fare jump out as the three likely predictors. Attempting to model survival using Sex, Pclass, and Fare in a logistic regression shows that both Sex and Pclass improve model fit, but Fare does not.

```
## Analysis of Deviance Table
##
## Model 1: Survived ~ Sex
## Model 2: Survived ~ Sex + Pclass
## Model 3: Survived ~ Sex + Pclass + Fare
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       889     917.80
## 2       888     827.20  1   90.608   <2e-16 ***
## 3       887     826.67  1    0.521   0.4704
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A likely explanation of this is the high correlation between Pclass and Fare. Fare doesn't ostensibly provide any insight not provided by Pclass.

```
## [1] -0.5494996
```

Low correlation across the entire sample does not necessarily mean a variable will not add value to the model. As shown below, there are several more terms that add significant value to the model.

```
## Analysis of Deviance Table
##
## Model 1: Survived ~ Sex
## Model 2: Survived ~ Sex + Pclass
## Model 3: Survived ~ Sex + Pclass + projAge
## Model 4: Survived ~ Sex + Pclass + projAge + SibSp
## Model 5: Survived ~ Sex + Pclass + projAge + SibSp + Sex:Pclass
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1       889     917.80
## 2       888     827.20  1   90.608 < 2.2e-16 ***
## 3       887     808.96  1   18.239 1.949e-05 ***
## 4       886     787.47  1   21.487 3.563e-06 ***
## 5       885     764.48  1   22.990 1.628e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

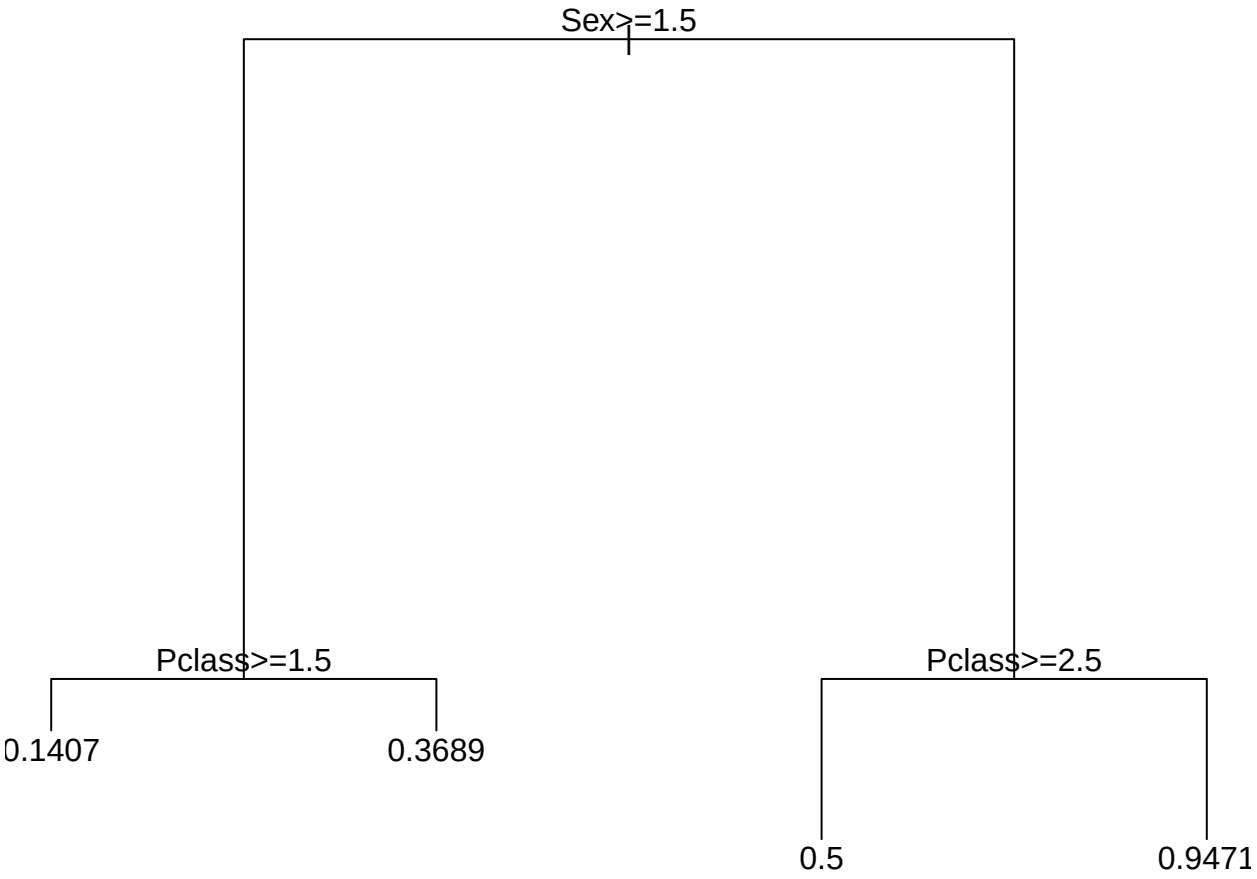
Secondary Predictors - Age and Siblings

How can we explain the effect of Age and SibSp on the model? Why is the interaction term between Sex and Pclass important?

To answer these questions, let's look at a decision tree that takes Sex and Pclass into account.

Here, we see that 94.7% of 1st and 2nd class women survived versus 50% of 3rd class women. Additionally, 14% of 2nd and 3rd class men survived versus 36.9% of 1st class men. Two things are clear from this tree:

- Pclass matters much more for women than for men, explaining the importance of the interaction term
- To improve the model, our focus should be on 3rd class women and 1st class men.



Let's subset the data into our new groups of interest and re-correlate with survival.

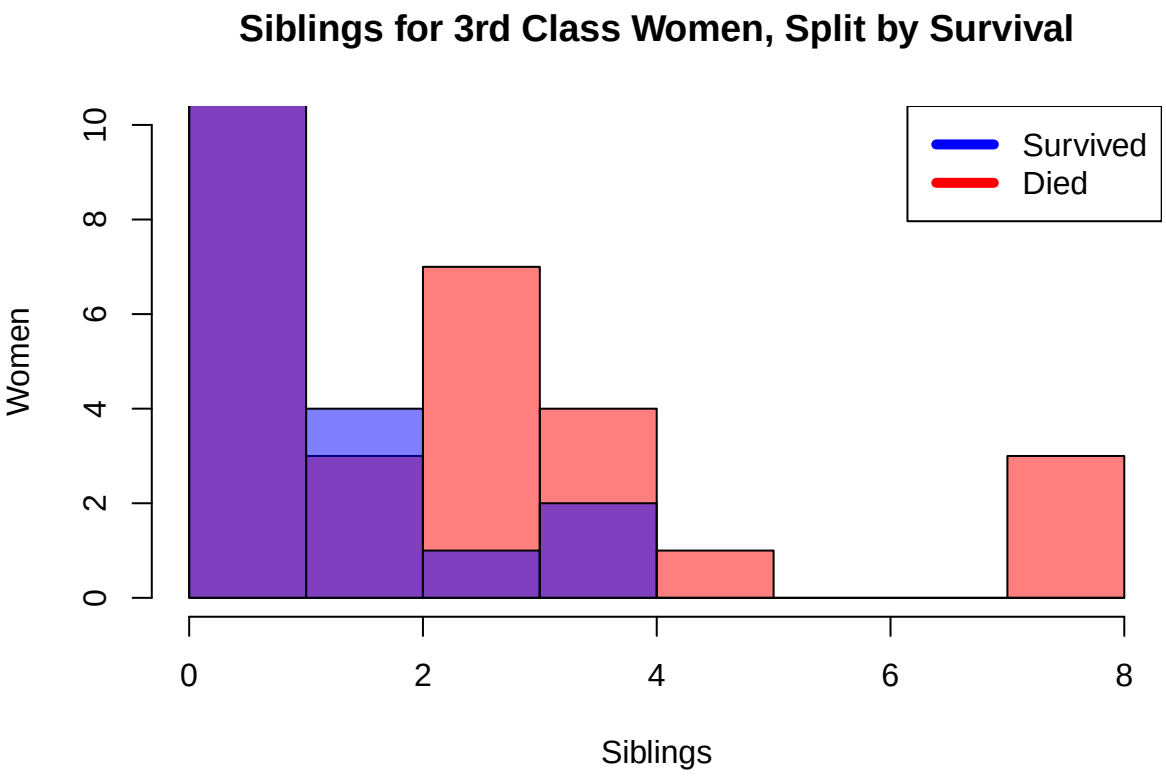
1st class men

```
##         Age     projAge        Fare       SibSp    Embarked       Parch
## -0.27056553 -0.27056553  0.09124571  0.06835598 -0.06544025  0.01529884
```

We can now more easily see why Age improved the model. Within the group of 1st class men, children are more likely to survive and older men are more likely to die.

3rd class women

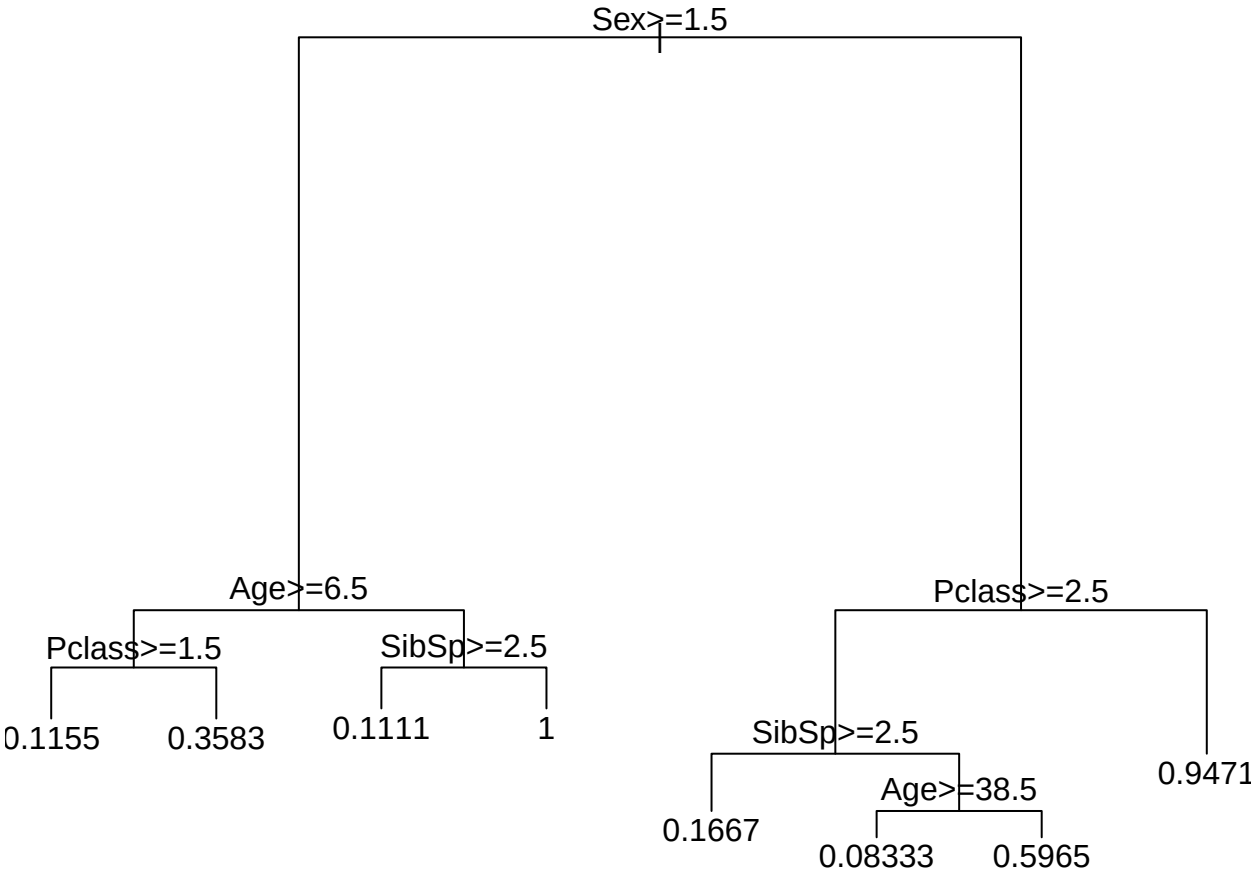
```
##       Fare   Embarked      SibSp        Age    projAge      Parch
## -0.3019067 -0.2017206 -0.1906067 -0.1766173 -0.1766173 -0.1700098
```



We can now more easily see why SibSp improved the model. Within the group of 3rd class women, those with fewer siblings are more likely to survive. There also does appear to be an effect from Fare and Embarked emerge in this subset. However, this investigation will not be pursued in this paper.

A (More) Complete Decision Tree

Let's add our secondary predictors to our decision tree:



Here, we can see the effect of Age on 1st class men and SibSp on 3rd class women. We can also see that in fact, both Age and SibSp have an effect on our other groups as well.

The large groups still with room for improvement at this point are shown below along with the group survival correlations.

- 1st class men over the age of 6.5

```
##       SibSp    Embarked        Fare       Parch
##  0.08748861 -0.06464291  0.06056810 -0.02093851
```

- 3rd class women under the age of 38.5 with 2 or fewer siblings

```
##    Embarked        Fare       Parch
## -0.19870102 -0.05232360 -0.01619802
```

There is no obvious improvement to be made for 1st class men over the age of 6.5.

It is possible that accounting for embarkment could improve predictions for 3rd class women under the age of 38.5 with 2 or fewer siblings. However, this investigation will not be pursued in this paper.

Logistic Regression, Decision Tree, or Random Forest?

Now that we know which variables we want to include, let's compare the relative success of these techniques on the train and test data sets.

Logistic Regression

Below is the proportion of correct responses in the train set using logistic regression on Sex+Pclass+SibSp+projAge+Sex:Pclass.

```
## [1] 0.8125701
```

On the test data, this performs at about .77 correct.

Logistic Model Tree

Below is the proportion of correct responses in the train set using a logistic model tree partitioned on Sex+Pclass+projAge and leaves modeled with Parch+SibSp+Embarked.

```
## [1] 0.8282828
```

On the test data, this performs at about .78 correct.

Conditional Tree

Below is the proportion of correct responses in the train set using a conditional inference tree on Sex+Pclass+SibSp+projAge.

```
## [1] 0.8395062
```

On the test data, this performs at about .75 correct.

Random forest

Below is the proportion of correct responses in the train set using a random forest on Sex+Pclass+SibSp+projAge.

```
## [1] 0.8193042
```

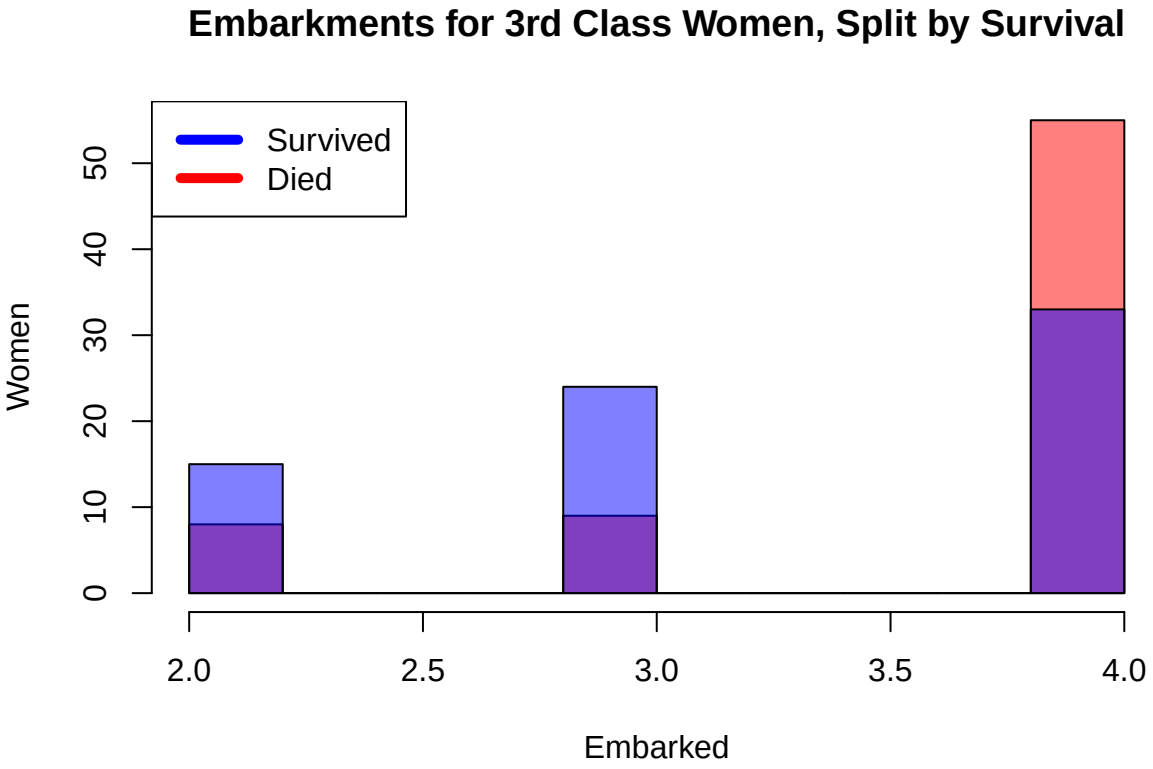
On the test data, this performs at about .75 correct.

Room for further improvement

Embarkment's role in survival was not explored in this analysis. It had a weak, but statisticaly significant correlation with Survival across the entire population, as well as with 3rd class women.

```
##         Sex      Pclass        Fare    Embarked       Parch         Age
## -0.53882559 -0.35965268  0.26818862 -0.18997851  0.09331701 -0.07722109
##     projAge       SibSp
## -0.07722109 -0.01735836
```

```
##       Fare   Embarked      SibSp        Age    projAge      Parch
## -0.3019067 -0.2017206 -0.1906067 -0.1766173 -0.1766173 -0.1700098
```



Embarkment also potentially improves upon the regression model and can't be explained away as redundant because it is not as strongly correlated with Pclass as Fare was.

```
## Analysis of Deviance Table
##
## Model 1: Survived ~ Sex
## Model 2: Survived ~ Sex + Pclass
## Model 3: Survived ~ Sex + Pclass + projAge
## Model 4: Survived ~ Sex + Pclass + projAge + SibSp
## Model 5: Survived ~ Sex + Pclass + projAge + SibSp + Sex:Pclass
## Model 6: Survived ~ Sex + Pclass + projAge + SibSp + Embarked + Sex:Pclass
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1       889     917.80
## 2       888     827.20  1   90.608 < 2.2e-16 ***
## 3       887     808.96  1   18.239 1.949e-05 ***
## 4       886     787.47  1   21.487 3.563e-06 ***
## 5       885     764.48  1   22.990 1.628e-06 ***
## 6       884     758.76  1    5.718   0.01679 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##        Fare      Pclass    Survived         Sex         Age     projAge
## -0.28679953  0.25329050 -0.18997851  0.11984079 -0.04483030 -0.04483030
##       SibSp       Parch
##  0.03762860  0.01627835
```

It also does potentially improve upon the regression model and isn't as strongly correlated with Pclass as Fare was.

Besides embarkment, certain variables were removed from the start. The marginal benefits of Name, Ticket, and Cabin were deemed too small to warrant a thoughtful translation of those fields for use in modeling. It is however always possible these fields do contain useful information which could improve the model.