Sebastian Kurpiel

Comp 379

9/9/17

Homework 1: Titanic Survival Prediction

Originally in the planning phase I wanted to use a decision tree to figure out the survival rate of a passenger. (I thought it would be interesting to see if I could include myself and see if I survived.) However when coding the tree I decided that it would be imprecise because moving along each child of the tree I would have to assign weights of survival and figuring out how to give an accurate weight was proving to be difficult. One difficulty was that women were the most likely to survive, 75% compared the men 19%[1] so figuring out that weight without skewing the data was proving tricky.

I soon decide to switch to a logistic regression[2] approach because it gives an output of 1 or 0 which I interpreted it as survived and dead. My implementation of logistic regression starts off by creating the shape of the dataset(line 32). Then it checks if there are any duplicates in the datasets(line 35) that could potential skew the data and line 43 checks if we a NaN, which is important because our dataset contains them, Female,Male,Embark port,etc. The next part of the code calculates the percentage of survival, it does so by taking the mean survival stats of each survivor. First class passengers (63%)had the greatest chance of survival while the second(47%) and third(24%) had the lowest. Female passengers (75%) had the highest chance of survival over men(19)%. Surprising me having fewer siblings also meant that you had a higher chance of survival, probably because it was typical for the wealthy to have less children and they tended to ride in first class. Supporting this was that a parch size of 3(includes mother, father, and child) had a higher rate of survival over ever other class. After this was done, I decided that certain information wasn't needed for calculating the survival;

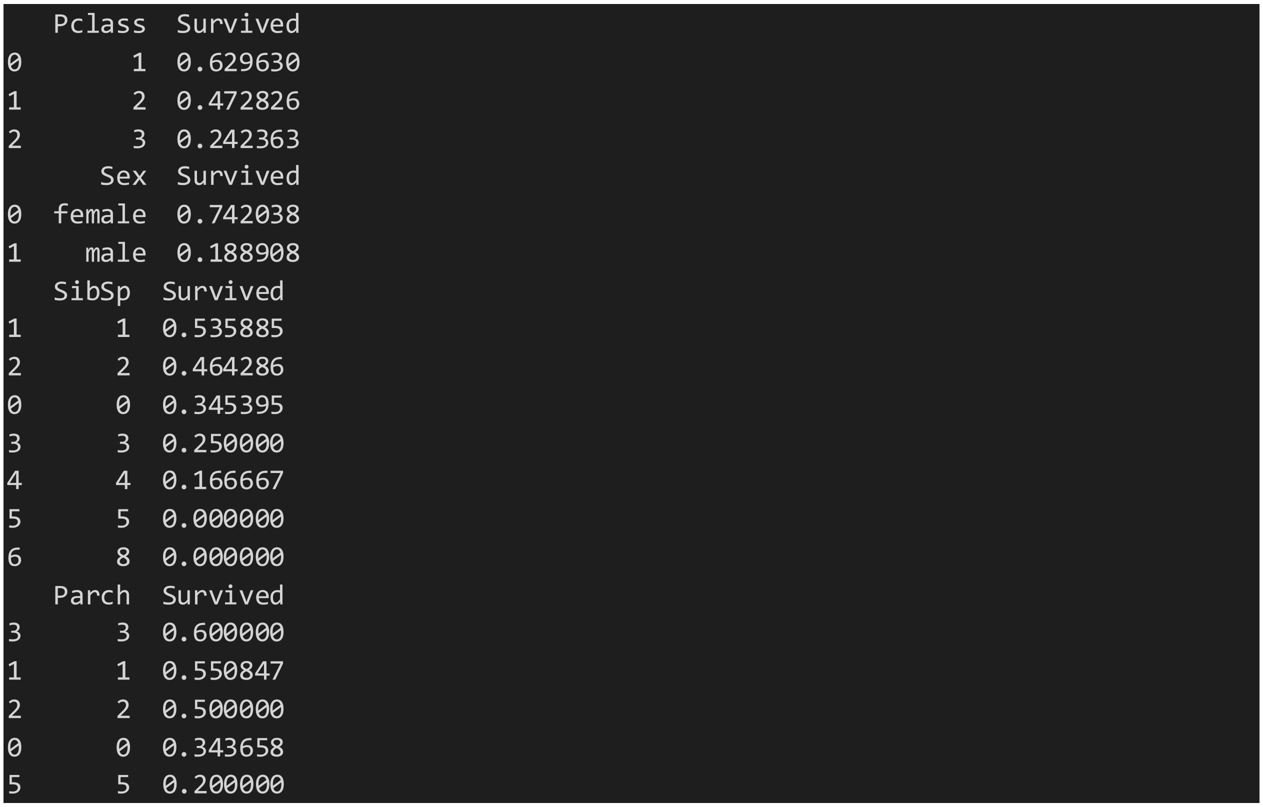
[1] Demographics: http://www.icyousee.org/titanic.html

[2] Where I found the information about Logistic Regression: https://medium.freecodecamp.org/the-hitchhikers-guide-to-machine-learning-algorithms-in-python-bfad66adb378

embarking port. I then added a couple of new features to the data such as if child and rich, if rich and a child you had a higher chance of survival. Another feature was title, I noticed that the better the title and the female specific title the higher chance of survival. The last feature I added was family size, which showed an advantage to survival if the family size was smaller. From this clean data I was ready to implement the logistic regression which to my surprise wasn't that difficult .

My code exports a .csv file called titanic results. I have only had time to spot check the results but it seems to be accurately giving results, I check by comparing passenger ids to names then I type in it in an online database[3] to confirm my results. Overall I think that logistic regression is the quickest way to calculate the odds of surviving the titanic, took 2.6 secs for the results. However, I will probably be checking if other machine learning algorithms do a better job.

Tables from the Python Output:



```
   Pclass  Survived
0       1  0.629630
1       2  0.472826
2       3  0.242363
      Sex  Survived
0  female  0.742038
1    male  0.188908
   SibSp  Survived
1      1  0.535885
2      2  0.464286
0      0  0.345395
3      3  0.250000
4      4  0.166667
5      5  0.000000
6      8  0.000000
   Parch  Survived
3      3  0.600000
1      1  0.550847
2      2  0.500000
0      0  0.343658
5      5  0.200000
```

[3] The Online database I used to check my results: https://www.encyclopedia-titanica.org/titanic-victims/

```
4      4  0.000000
6      6  0.000000
```