

Titanic: Machine Learning from Disaster

EECS 510

Xiaoyang TAN
Zhaoyang LIU

May 20, 2014

1 Current work

1.1 Defining the problem

Information is given on a training set of passengers of the Titanic, for which the survival outcome is known. Given the training set information, the goal remains to predict each passenger's survival outcome from a test set of passengers. The details of the challenge are given on the Kaggle site.

We will apply machine learning tools to solve the problem. In this case, we may consider approaches based on random forests. And will use Python package to plot the result. The python package used includes:

- NumPy
- Pandas
- SciKit-Learn
- SciPy
- StatsModels
- Patsy
- Matplotlib

1.2 Analyze the data

In the training dataset, there are 891 passengers. Each passenger has 12 attributes. Except the "passengerId" and "Survived" attribute, we need to consider 10 other attributes and predict the survival possibility.

1.2.1 Take care of missing values

The features *Ticket* and *Cabin* have many missing values and so can't add much value to our analysis. To handle this we will drop them from the data frame to preserve the integrity of our dataset. What's more, we will remove NaN values from every remaining column. Using **drop()** and **dropna()** function in could easily achieve goal. Now we have a clean and tidy dataset that is ready for analysis, we cut the dataset from 891 to 712, and get 8 effective attributes to do prediction.

1.2.2 Graphical view of data

The point of this competition is to predict if an individual will survive based on the features in the data like:

- Traveling Class (called pclass in the data)
- Sex
- Age

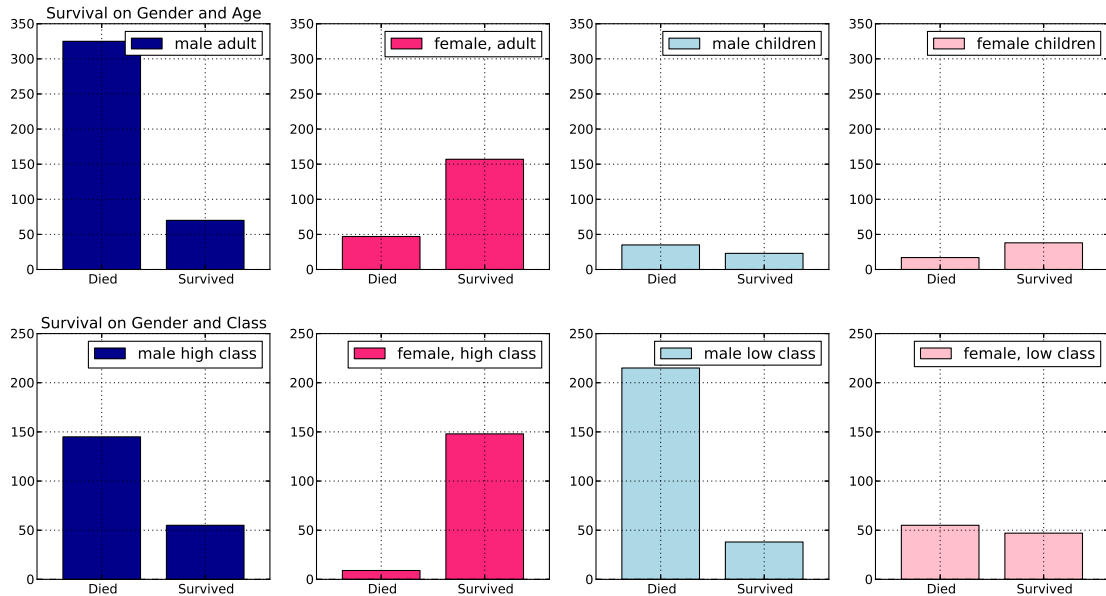


Figure 1: Survival on gender, age and class

Figure 1 shows the distribution of survival on gender age and class. *****

1.3 Simple trying

In the film, they tried to rescue women and children first, so a good guess would be on gender. From the graphical view of the data, its clear that although more men died and survived in raw value counts, females had a greater survival rate proportionally(25%), than men (20%).

After applying that all the female would survive, the survival rate become 36.36%, which comes close to the original rate 38.38%. However, we can refine our results by considering other variables.

We ruled out the age factor, as the survival rate doesn't change much with or without the involvement of age. Thus we split the fare class into four payments range and assumes that any group with more than half survivors will always be modeled to survive, while the rest will be mapped to death. The result is different from the previous one in that women in 3rd class who paid more than \$20 will not survive, which bring the survival rate closer to training set.

1.4 Supervised machine learning

2 Next step

2.1 SVM

2.2 Random Forest

3 Possible results