Titanic survival evaluation

# ****Introduction/Business Problem.****

The problem is simple: to use machine learning to create a model that predicts which passengers survived the Titanic shipwreck based on the information provided about the passengers. This is not a secret that first class passengers were the ones who mostly survived, but the idea is to find more factors (other than the class of the ticket) that could influence the survival of a given passenger.

# ****Data.****

The data is gathered in a csv file with 891 entry. Each entry represents a passenger and consists of his/her name, age, marital status, ticket class, sex, age in years, no. of siblings/spouses aboard the Titanic, no. of parents/children aboard the Titanic, ticket number, passenger fare, cabin number and the port of embarkation. And of course, either a person survived or not.

The first five rows of the file are these:

|  |
| --- |
| PassengerId,Survived,Pclass,Name,Sex,Age,SibSp,Parch,Ticket,Fare,Cabin,Embarked |
| 1,0,3,"Braund, Mr. Owen Harris",male,22,1,0,A/5 21171,7.25,,S |
| 2,1,1,"Cumings, Mrs. John Bradley (Florence Briggs Thayer)",female,38,1,0,PC 17599,71.2833,C85,C |
| 3,1,3,"Heikkinen, Miss. Laina",female,26,0,0,STON/O2. 3101282,7.925,,S |
| 4,1,1,"Futrelle, Mrs. Jacques Heath (Lily May Peel)",female,35,1,0,113803,53.1,C123,S |

It is already understandable from these four entries that the data is incomplete somewhere. So, the whole dataset should be checked for consistency, the second column (Survived) should be moved to become the last column and several columns (such as Name, Ticket, Fare, Cabin, Embarked) removed from the trained model.

# **Methodology.**

As the we have to divide the passengers into two groups and we already know which groups that should be (survived or not), it is definitely the problem of classification.

First of all, I tried to figure out what data is not significant for the result. I dropped the rows with passengers’ names, ticket number, cabin number, the port where they embarked the ship and the price for the ticket.

I supposed the class of the ticket, the gender and the age very important. I also considered that the fact if a person was travelling with some kind of family members could influence the result.

To understand which method suits better I divided the train set into two parts, to train and to test using different methods and to understand which algorithm suits better.

After I understood that KNN algorithm gives the best result, I took the whole train set and trained the model, then applied the model to the test set.

# **Results.**

The results as expected gave me the view that among all that managed to survive nearly half were the passengers of the first class. Another half consisted of the 2nd and 3rd class passengers almost evenly. But as the ship contained much more passengers of the 3rd class than of the 2nd and of course of the 1st, we can see that more than half of the passengers of the 1st class survived, one third of the 2nd class and only every fifth passenger of the 3rd class.

# **Discussion.**

An interesting observation about the sex of the passengers. There were slightly more women than men in the first class and major part of them survived. As for men, only every third man in the first class survived. As there were enough dinghies to hold the first class, I supposed they just gave their place to the ladies and children of all the classes. That can be confirmed by the observations of the second class, where the rate of survived women is very high, while probably all the men died.

# **Conclusion.**

I know that behind all of these numbers are the real people’s lives, I feel really sorry for them. But this tragedy taught us a lot: how safety is not something to play with, how important it is to be ready for the most improbable thing. From the statistical point of view this is a great opportunity to study the distribution of the results, and to see which factors, that may not have been obvious from the first view, can influence the result.