Prediction of Survivals on the Titanic

LI, JUNRONG
ZHANG, WENHAO
LIU, CHENG
Hong Kong University of Science and Technology
April 13, 2018

Abstract

The sinking of the RMS titanic is one of the most severe shipwrecks in history, which directly leads to the death of over 1,500 people due to the limited number of lifeboats. Among those escaping from death, some groups of people, such as women, children and upper-class, were more likely to survive for certain reasons. This essay is intended to unearth the characteristics of surviving passengers. At the same time, we also want to find out which prediction model outperforms others in this case.
Based on the given training and testing set of passenger information, we conduct complement on the missing cells, cleansing and reconstruction on both datasets. Under careful investigation, we choose "Age", "Embarkation", "Sex", "Fare", "Passenger Class" and "Family Size" as the inputs which may greatly affect the survival possibility of the passenger. This dataset features small size, weak linear relationship among inputs and potential hidden logic layers. Several classification methods are adopted to predict the characteristics, including SVM, XGBoost, KNN, SGD and so on.
The research outcome shows that SVM gives the best accuracy of 78.947%, slightly higher than XGBoost. ADABoost and Decision Tree are also useful at predicting the survival of passengers in this severe disaster, which, however, may be more meaningful or insightful in other cases.

1 Introduction

The sinking of the RMS Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. This sensational tragedy shocked the international community and led to better safety regulations for ships.
One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others, such as women, children, and the upper-class.primis in faucibus.
Given a data set of 1600 passengers in this problem, we are going to analysis the significance of features, extract useful information, and then construct several training models

based on different combinations of features and weights. By comparison, the model with highest prediction accuracy in the testing set would come out as a solution.

2 Methodology

2.1 Data Acquring

The Python Pandas packages helps us work with our datasets. We start by acquiring the training and testing datasets into Pandas DataFrames. We also combine these datasets to run certain operations on both datasets together.
The structure of the extracted data is shown in Figure 1.

Data Dictionary

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

Figure 1: Data Dictionary

2.2 Data Analysis

2.2.1 General

What we can use to predict are all features except "survival" and "id" to predict the survival situation.

The given dataset contains two parts, one for training and the other for testing. The two set embodies 819 and 418 entries correspondingly. Each piece of data uses 12 features to describe a single passenger including "PassengerID", "Name", "Survival status", "Passenger class", "Sex", " Age", "Number of siblings and spouses abroad the Titanic", "Number of parents and children aboard the Titanic", "Ticket Number", "Passenger Fare", "Cabin number" and "Port of embarkation". We can use all features except "survival" and "id" to predict the survival situation. Among these features, "PassengerID", "Name", "Ticket number" and "Cabin number" are text strings varying from person to person. "Survival status", "Sex" and "Pork of embarkation" are represented in ordinal number or texts. The rest features are in the form of discrete or continuous numbers.

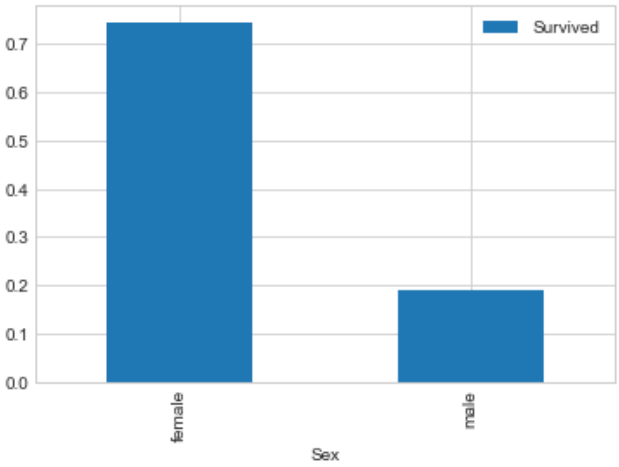

Figure 2: survived vs sex

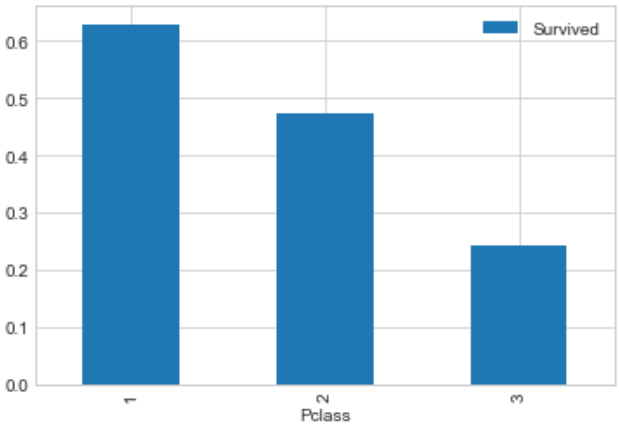

Figure 3: survived vs pclass

2.2.2 Description

Passenger ID - Each passenger owns a unique ID.
Name - Each passenger owns a unique piece of text, which embodies his title, family name and last name.
Survival Status - "1" stands for survival and "0" stands for no survival.
Passenger class - Ordinal number. Higher number indicates less advanced cabins.
Sex - "1" stands for male and "0" stands for female
Age - Discrete cardinal number
Number of siblings and spouses abroad the Titanic - Discrete cardinal number
Number of parents and children aboard the Titanic - Discrete cardinal number
Ticket Number - String
Passenger Fare - Continuous cardinal number
Cabin number - Each passenger owns a unique cabin number.
Port of embarkation - A single letter is used to indicate the passenger's departure port. C = Cherbourg, Q = Queenstown, S = Southampton

After basic investigation on data, we observe several indicative facts:

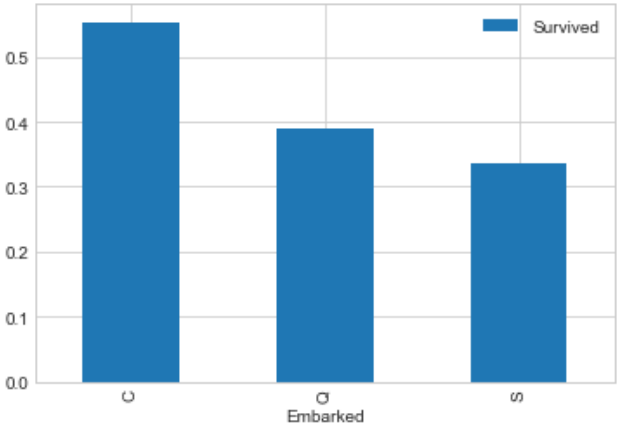

Figure 4: survived vs embarked

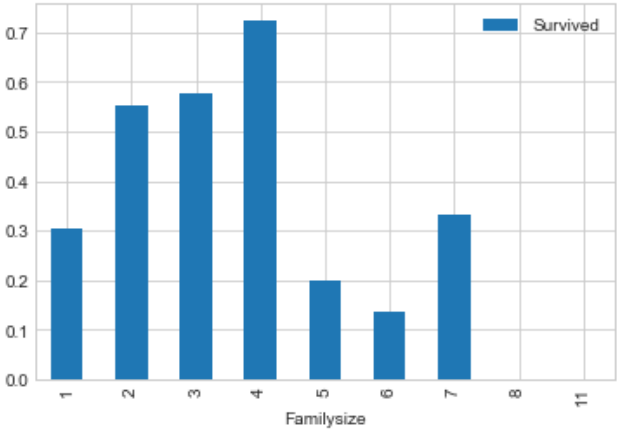

Figure 5: survived vs family size

1. Gender
   Figure 2 shows that among all the survived people, females accounted for more than 70%. We believe that this is due to the universal awareness that women enjoyed priority to get aboard on the lifeboats, which greatly increased their possibility of survival.

2. Cabins
   Figure 3 shows that passenger living in more advanced cabins were more likely to survive. The investigation afterwards shows that the hull collided with the icebergs at the lower part. As more advanced cabins were located in the higher floors of the cruise ship and also further away from the collision part, passengers there got more time for escape, which increases their possibility of survival. Embarkation

3. Embarkation
   Figure 4 shows that passengers departing from Cherbourg were more likely to survive, while those from Southampton were less likely.

4. Family Size
   Figure 5 shows that passengers from middle-sized family (which embodies siblings, spouses, parents and children) consisted of 4 people were more likely to survive

2.2.3 Data Cleansing

The original data has some shortcomings and null values. If we use it to do the prediction, it might have a negative influence on the result of learning. So we need to do some pre-processing on the data.

Following are the actions we took on some of the features:

Age: We use mean value to fill in the N/A values in the data set.
Fare: We use median value to fill in the N/A values in the data set.
Parch: We use mean value to fill in the N/A values in the data set.
Embarked: We use the port with highest frequency to fill in the N/A values in the data set.
Pclass: We use the class with the highest frequency to fill in the N/A values in the data set.
SibSp: We use mean value to fill in the N/A values in the data set.

For Embarked and Pclass, we use the one with the highest frequency because they are more general. And for string type of features, we cannot generate mean or median value. And we cannot easily determine the port where a passenger get on, thus it is hard to use a specific value for that attribute. And if we use the one with the highest frequency, it avoids being a noise to a smaller group of data. Since if we predict it wrongly, it causes a bias.

3 Result and Discussion

Table 1: Accuracy of Different Methods

Method	Measure Accuracy
SVM	0.78947
XGBoost	0.76555
ADABoost	0.76555
Random Forest	0.75119
Decision Tree	0.71770
Bagging	0.69700
5-layer Neural Network	0.68899
KNN	0.65555
SGD	0.62679

The SVM model gives the highest accuracy of 78.95%, slightly higher than XGBoost ?s 76.56%. After data processing, we only select 6 features for prediction. In SVM model, the algorithm is going to increment the number of dimensions of data, which somehow increases the precision of the prediction process. On the other hand, due to the limited number of features provided, the XGBoost model cannot generate enough subtrees to wipe out the bias created by the correlation among different inputs. And thus the accuracy is reduced.

KNN model only generates an accuracy of 65.56% in this case. The main reason is the limited number of features adopted, which lowers the precision of clustering. Another one is that although two individuals may share similar characteristics in KNN model, in reality, they may be faced with quite different situations that may lead to various survival situations, meaning that similar characteristics don?t indicate similar result, which decreases the prediction accuracy.

SGD is a very scientific and serious method. It highlights the compulsory relationship between the features and the results. Then it can gradually reduce the lost function and approaches to the best model. However, there is not an obvious and true mathematical relationship that can absolutely determines the survival situation of a passenger. Even if two passengers are very similar, the actual relationship can vary and we cannot say every-

thing in 100%. So this kind of hard-margined classification method will perform badly on this dataset. NN focuses on the relationship between hundreds of features and then we can build up the relationship between features to predict the result. It can have a great performance in linking many features and highlights the concentrate points as well as the grouping. However, there are only 11 features available here and I only used around 6 of them. Which is a very small number, then the neural network turns out to be meaningless for this kind of prediction problem.

4 Conclusion

This is an interesting data set for getting familiar with Kaggle. We learned how to use pandas library to frame and wrap up the data. We used matplotlib to visualize the features and figure out the potential relationship between them. Then we can choose the features that are useful. For the machine learning parts, we used xgboost and sklearn to make use of the powerful implementations of various algorithms.

The data here is fun, with a lot of information hidden under the cover. There indeed are relations between features of this dataset but we found that the relations are not that serious and mathematical. It might be hard to approach them using pure mathematical methods. We need to add more considerations from human points of view. And this is the point why this dataset is interesting and worthwhile working on.

APPENDIX

Running environment:
Python 3.6
Library: sklearn, matplotlib, pandas, numpy
README:
There are 2 python files included in this project.
Titanic_pred.py: this includes the file handling and all the models for machine learning.
Plots.py: this file plots the data relationship required for data visualization.