Titanic dataset analysis

Bogachev Aleksei

August 24, 2022

Abstract

Data analysis and models for the legendary machine learning competition on Kaggle.

Contents

Chapter 1

Introduction

Perhaps, the sinking of the RMS Titanic is the most infamous shipwreck in history. According to the Wikipedia, the RMS Titanic was the largest ocean liner in service at the time. It had advanced safety features, such as watertight compartments and remotely activated watertight doors. The ship was widely considered "unsinkable". However, the Titanic sank in the early morning of 15 April 1912 in the North Atalntic Ocean during her maiden voyage from Southampton to New York City. There were an estimated 2224 people on board when the ship collided with an iceberg [4],[3].

In accordance with existing practice, Titanic's lifeboat system was designed to ferry passengers to nearby rescue vessels, not to hold everyone on board simultaneously; therefore, with the ship sinking rapidly (the ship had sank in 2 hours and 40 minutes) and help still hours away, there was no safe refuge for many of the passengers and crew with only 20 lifeboats. Poor management of the evacuation meant many boats were launched before they were completely full [3].

The shipwreck resulted in the deaths of more than 1500 people, makng it one of the deadliest in history [3].

Without a doubt, there was an element of luck involved in surviving, but, possibly, some groups of people were more likely to survive than others. The Titanic ML competition on Kaggle offers participants to predict which of the passengers survived the shipwreck using passenger data[5].

In this report I'm going to describe my solution of the Titanic ML competition's task. My workflow will be based mostly on the "Machine Learning project checklist" from the book [1]. I really appreciate this book and highly recommend reading it to anyone starting to learn about machine learning.

Dozens of articles dedicated to this competition and hundreds of solutions of this task are available in the Internet. Therefore, I won't cite to all materials seen, but I'll try to give several useful refrences.

Chapter 2

Task Description

In this section the task is described according to "Machine Learning project checklist" [1].

2.1 Goal

To predict if a passanger survived the sinking of the Titanic or not.

2.2 Crrent Solutions

There are dosens of solutions available on the discussion forum and on the Internet.

2.3 Frame the Problem

- Supervised learning
- Classification
- Binary classification (survived of not)
- Batch learning (no continuous flow of data and the dataset is small)

2.4 Performance metrics

This competition evaluates the **percentage of correctly predicted passengers** (accuracy).

There are also several useful metrics for evaluating the performance of a classification system:

- precision,

- recall,
- F_1 score,
- precision/recall curve,
- ROC curve,
- ROC AUC score.

2.5 Target performance

The leaderboard of this competition contains almost 14000 entries. It's available in the form of the csv-file. An excerpt from the leaderboard is presented in the table 2.1.

Table 2.1: Excerpt from leaderboard

TeamId	TeamName	SubmissionDate	Score
6987444	no name	2022-08-23 18:16:28	1.0
720238	rosh	2022-06-26 10:58:42	1.0
8814675	nikolai otvetchikov #2	2022-06-26 13:59:39	1.0
8821160	Vibhav Rathkanthiwar	2022-06-26 15:28:12	1.0
6590016	Osman Altuntas	2022-07-24 15:40:15	1.0

The descriptive statistics is shown in the table 2.2.

Table 2.2: Descriptive statistics of scores

Statistics	Value
count	13915.000000
mean	0.760751
std	0.075145
min	0.000000
25%	0.765550
50%	0.775110
75%	0.777510
max	1.000000

The median score is about 0.775, but less than 3% of the solutions have a score above 0.8. Thus, **an accuracy score equal to or greater than 0.8 would be a very good result**. Figure 2.1 shows ECDF of the scores in the leaderboard. In this figure, the red lines mark the score 0.8 and the corresponding proportion of solutions.

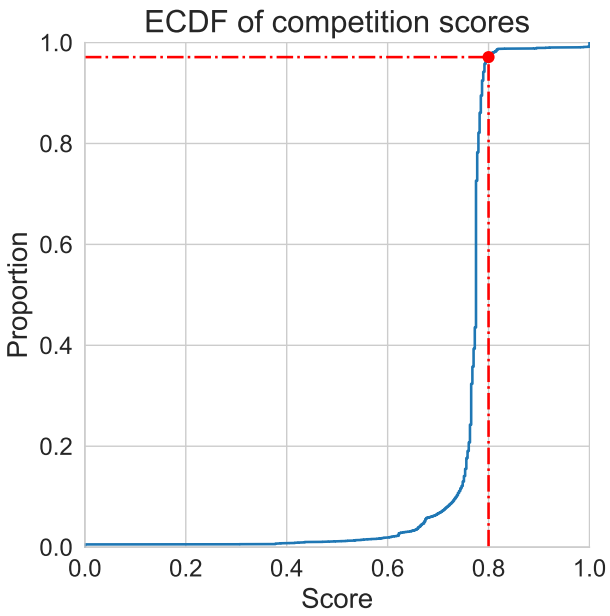

Figure 2.1: Leaderboard Scores ECDF

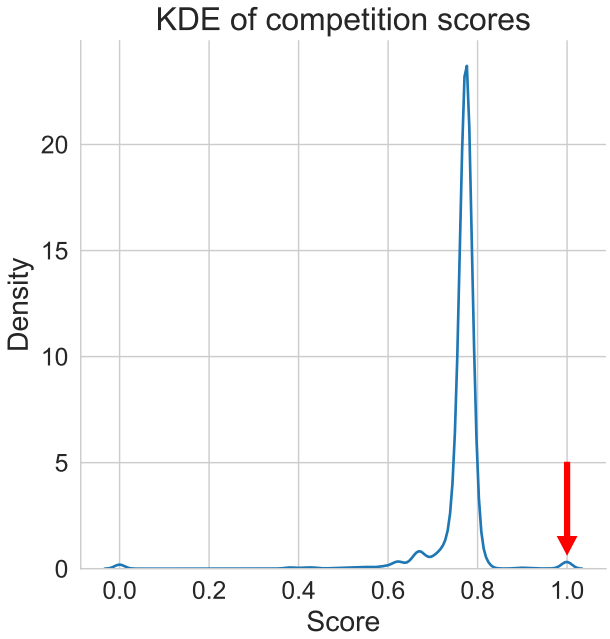

Figure 2.2: Leaderboard Scores KDE

There are several solutions with a score equal to 1.0. These solutions are marked with a red arrow in the figure 2.2. Have authors reached perfection?

I guess, this solutions appears, because there is an exact solution on GitHub. Possibly it is the data extracted from Encyclopedia Titanica[6] or from OpenML. Some authors in their notebooks honestly warn other users about the existence of such a possibility, for example, this one [2].

2.6 Data Dictionary

1. **PassengerId** – Passenger ID.
2. **Survived** – Survival:
    - 0 = No,
    - 1 = Yes.
3. **Pclass** – Ticket class:
    - 1 = 1st,
    - 2 = 2nd,
    - 3 = 3rd.
4. **Name** – Passanger's name, for example, "Braund, Mr. Owen Harris".
5. **Sex** – Gender:
    - male,
    - female.
6. **Age** – Age in years, for example 38.0.
7. **SibSp** – Number of siblings or spouses aboard the Titanic.
8. **Parch** – Number of parents or children aboard the Titanic.
9. **Ticket** – Ticket number, for example, A/5 21171.
10. **Fare** – Passenger fare, for ecample, 71.2833.
11. **Cabin** – Cabin number, for example, C85.
12. **Embarked** – Port of Embarkation:
    - C = Cherbourg,
    - Q = Queenstown,
    - S = Southampton.

2.6.1 Features

PassengerId, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, Embarked

2.6.2 Target

Survived

2.6.3 Variable Notes

- **pclass**: socio-economic status
  - 1st = Upper
  - 2nd = Middle
  - 3rd = Lower
- **age**: Age is fractional if less than 1. If the age is estimated, is it in the form of xx.5
- **sibsp** number of sibling/spouses aboard the Titanic
  - sibling = brother, sister, stepbrother, stepsister
  - spouse = husband, wife (mistresses and fiancés were ignored)
- **parch** number of parents (mother, father)/children (daughter, son, stepdauter, stepson) aboard the Titanic. Some children travelled only with a nanny, therefore parch=0 for them.

2.7 File Paths

- **training set**: ../datasets/train.csv
- **test set**: ../datasets/test.csv
- **example of a submission file**: ../datasets/gender_submission.csv

2.8 Assumptions

Women were more likely to survive than men.

References

[1] Aurelien Geron. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. 2nd. O'Reilly Media, Inc., 2019. ISBN: 1492032646.

[2] *How to be a Top LB Explained for Beginners*. URL: https://www.kaggle.com/code/suzukifelipe/how-to-be-a-top-lb-explained-for-beginners/notebook?scriptVersionId=99817039 (visited on 08/24/2022).

[3] *Sinking of the Titanic*. URL: https://en.wikipedia.org/wiki/Sinking_of_the_Titanic (visited on 08/24/2022).

[4] *Titanic*. URL: https://en.wikipedia.org/wiki/Titanic (visited on 08/24/2022).

[5] *Titanic - Machine Learning from Disaster*. URL: https://www.kaggle.com/c/titanic (visited on 08/24/2022).

[6] *Titanic Survivors - Names of all passengers and crew that survived. Complete list of Titanic survivors*. URL: https://www.encyclopedia-titanica.org/titanic-survivors/ (visited on 08/24/2022).