Titanic dataset analysis

Bogachev Aleksei

September 2, 2022

Abstract

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

Contents

1 Introduction 3

2 Task Description 4
- 2.1 Goal . . . 4
- 2.2 Crrent Solutions . . . 4
- 2.3 Frame the Problem . . . 4
- 2.4 Performance metrics . . . 4
- 2.5 Target performance . . . 5
- 2.6 Data Dictionary . . . 7
  - 2.6.1 Features . . . 7
  - 2.6.2 Target . . . 8
  - 2.6.3 Variable Notes . . . 8
- 2.7 File Paths . . . 8
- 2.8 Assumptions . . . 8

3 Preliminary Analysis 9
- 3.1 Shape of the dataset . . . 9
- 3.2 First rows of the dataset . . . 9
- 3.3 Data types and missing values . . . 9
- 3.4 Number of unique values . . . 10
- 3.5 Summary statistics . . . 11

4 Sample a Test Set 12

5 Exploratory Analysis 14
- 5.1 Pclass . . . 15
- 5.2 Name . . . 16
- 5.3 Sex . . . 17
- 5.4 Age . . . 18
- 5.5 SibSp . . . 18
- 5.6 Parch . . . 18
- 5.7 Ticket . . . 18
- 5.8 Fare . . . 18
- 5.9 Cabin . . . 18

5.10 Embarked 18

References **19**

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 the 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.

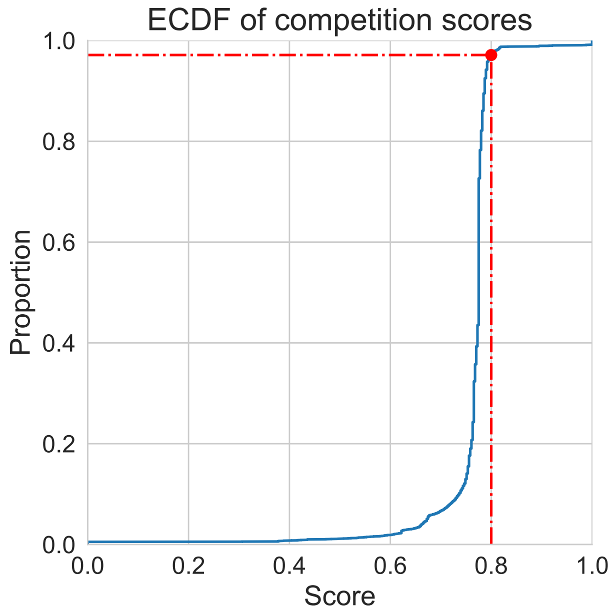

Figure 2.1: Leaderboard Scores ECDF

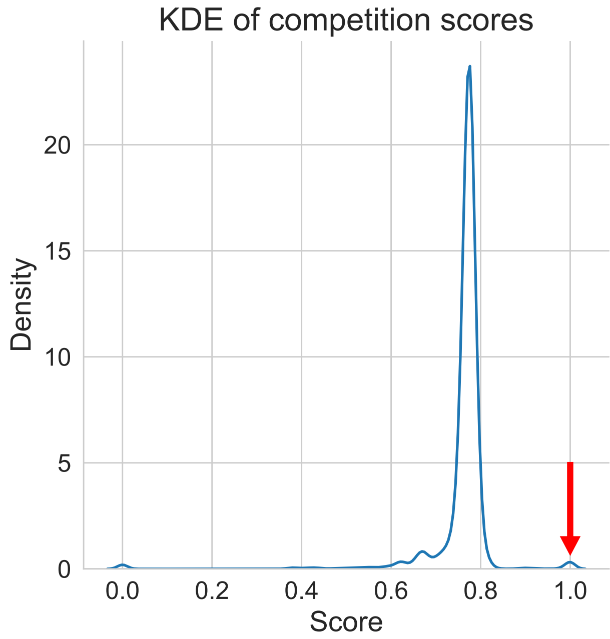

Figure 2.2: Leaderboard Scores KDE

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.

Chapter 3

Preliminary Analysis

3.1 Shape of the dataset

The dataset contains:

- 891 rows,
- 12 columns.

3.2 First rows of the dataset

An excerpt from the dataset is presented in the table 3.1.

Table 3.1: Excerpt from the dataset

	PassengerId	**Survived**	**Pclass**	**Name**	**Sex**	**Age**	**SibSp**	**Parch**	**Ticket**	**Fare**	**Cabin**	**Embarked**
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

The **"PassengerId"** feature is the ID of the passanger. It won't help in the analysis and will be dropped. Also, there are several missing values, and some values are categorical, for example, **"Pclass"** and **"Sex"**.

3.3 Data types and missing values

Table 3.2 contains types of the data in each column and numbers of non-null values. Table 3.3 contains numbers of missing values in each column.

Table 3.2: Data types and non-null counts

#	Column	Non-Null Count	Dtype
0	PassengerId	891 non-null	int64
1	Survived	891 non-null	int64
2	Pclass	891 non-null	int64
3	Name	891 non-null	object
4	Sex	891 non-null	object
5	Age	714 non-null	float64
6	SibSp	891 non-null	int64
7	Parch	891 non-null	int64
8	Ticket	891 non-null	object
9	Fare	891 non-null	float64
10	Cabin	204 non-null	object
11	Embarked	889 non-null	object

Table 3.3: Number of missing values in each column

#	Column	Number of missing values
0	PassengerId	0
1	Survived	0
2	Pclass	0
3	Name	0
4	Sex	0
5	**Age**	**177**
6	SibSp	0
7	Parch	0
8	Ticket	0
9	Fare	0
10	**Cabin**	**687**
11	**Embarked**	**2**

3.4 Number of unique values

Table 3.4 contains numbers of unique values in each column.

There are high-cardinality features with *object* dtype:

- Name
- Ticket
- Cabin
- PassengerId

Table 3.4: Number of unique values in each column

Column	Number of unique values	Column	Number of unique values
Name	891	Survived	2
Sex	2	Pclass	3
Ticket	681	Age	88
Cabin	147	SibSp	7
Embarked	3	Parch	7
PassengerId	891	Fare	248

This features, possibly, will need special preprocessing. Earlier, I noticed that the **"PassengerId"** feature is the ID of the passanger. It won't help in the analysis and will be dropped. Features **"Age"** and **"Fare"** are continuous.

3.5 Summary statistics

In this section, summary statistics for numerical and non-numerical attributes are presented in tables 3.5 and 3.6 respectively.

Table 3.5: Summary statistics for numerical attributes

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

Table 3.6: Summary statistics for non-numerical attributes

	Name	Sex	Ticket	Cabin	Embarked
count	891	891	891	204	889
unique	891	2	681	147	3
top	Braund, Mr. Owen Harris	male	347082	B96 B98	S
freq	1	577	7	4	644

Chapter 4

Sample a Test Set

The test set will be used to evaluate performance of a very final model and forecast the score in the competitions leaderboard. It may seems like it's to early to create a test set, but I'll do it to prevent data snooping.

I'm going to do stratified sampling with scikit-learn's StraifiedShuffleSplit to maintain equal ratio of men and women in the train set and the test set. Women seem to have had a better chance of surviving due to the "women and children first" protocol for loading lifeboats.

First, let's check how many passengers survived. Figure 4.1 illustrates these numbers. There are 342 (38.38%) survived passengers and 549 (61.62%) drowned passengers in the dataset, so the dataset is a bit skewed. However, it's most likely, there will be enough representatives of both classes in the test set.

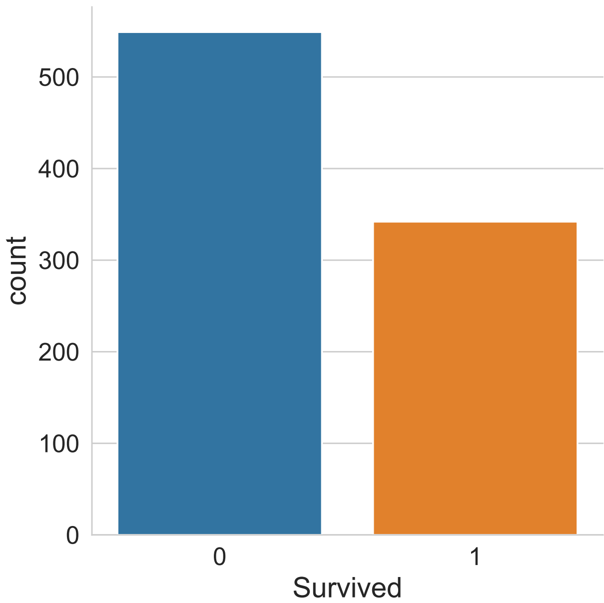

Figure 4.1: Number of survived and drowned passangers in whole dataset

Next, let's check the proportion of women among all survivors (figure 4.2), and check the proportion of survived women in each **"Pclass"** (figure 4.3).

Figures 4.2 and 4.3 show that in the entire dataset and in each **"Pclass"**,

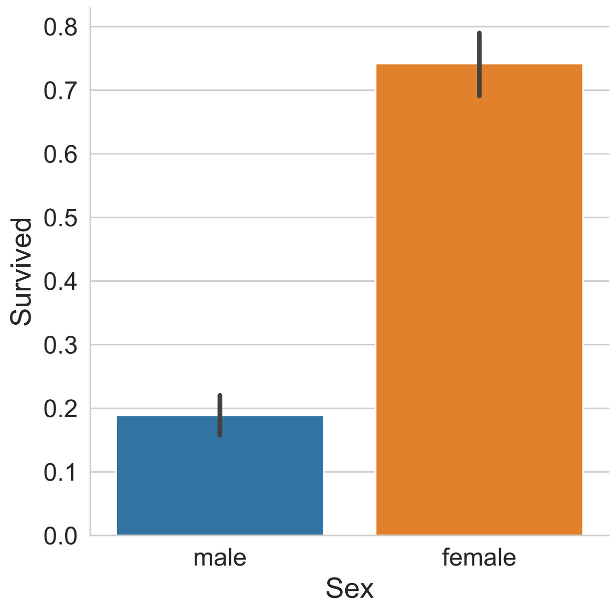

Figure 4.2: Proportions of survived men and women

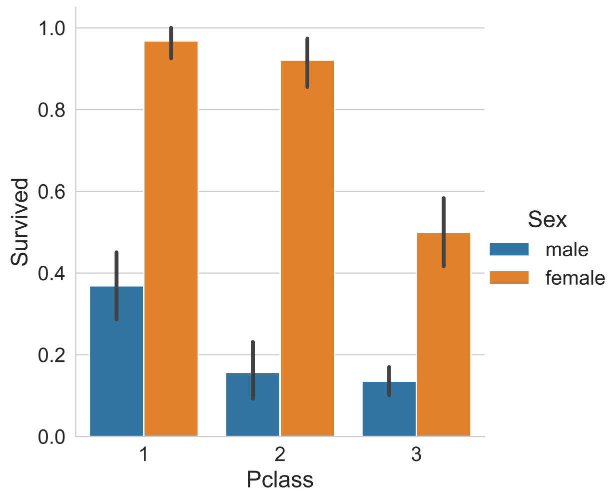

Figure 4.3: Proporotion of survived women and men in each **"Pclass"**

there are more female survivors than males. Thus it is reasonable to do stratification based on the passenger's gender. I will use 80% of the data for training and hold out 20% for testing, refer to the Jupyter Notebook for details.

Chapter 5

Exploratory Analysis

In this chapter, I will copy the training set to an exploratory set and work with it, because the exploratory analysis may require some data transformations, however I want to keep the training set untouched.

I will explore each attribute and its characteristics. Here is a list of dataset attributes:

- PassengerId
- Survived
- Pclass (page 15)
- Name (page 16)
- Sex (page 17)
- Age (page 18)
- SibSp (page 18)
- Parch (page 18)
- Ticket (page 18)
- Fare (page 18)
- Cabin (page 18)
- Embarked (page 18)

I won't consider **"PassengerId"** and **"Survived"** attribures. **"PassengerId"** attribute contains 891 passenger's IDs (see section First rows of the dataset and table 3.4), which won't help in building a model. **"Survived"** attribute is a target. A passenger survived if their **"Survived"** attribute is 1, else the passenger drowned (**"Survived"** attribute is 0). Figure 4.1 shows numbers of survived and drowned passengers in whole dataset.

5.1 Pclass

The **Pclass** attribute contains information about socio-economic status of the passenger:

- 1st = Upper
- 2nd = Middle
- 3rd = Lower

Let's estimate the number of passengers of each class in the exploratory set. The figure 5.1 illustrates this estimation.

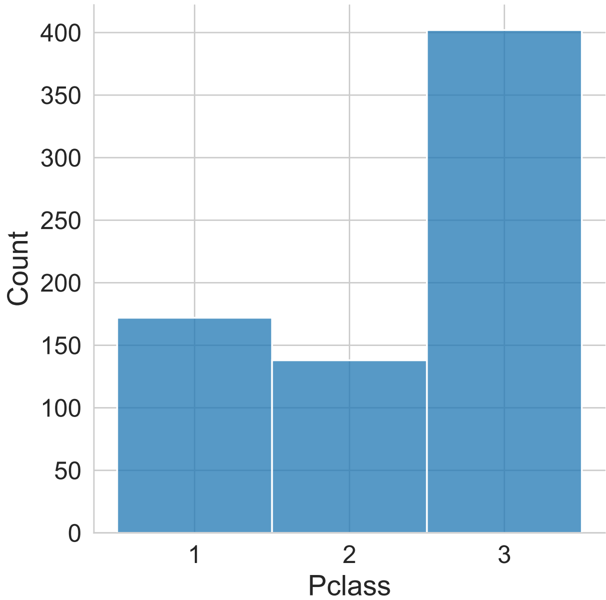

Figure 5.1: Number of passengers in each **Pclass**

The figure 5.2 shows the proportions of survived passengers for each **Pclass** in the exploratory set.

It looks like there were more passengers of the lower socio-economic class (**Pclass**=3), but they had less chance of surviving. **Pclass** is a class of the ticket, so it contains information about the location of the passenger's cabin. It is known that the cabins of passengers with low socio-economic status were located on lower decks, that is, further from the lifeboats, this explains why there are fewer survivors among them [4].

Finally, let's check the proportion of women among the survivors of each **Pclass**. Figure 5.3 show these proportions.

More women than men survived in each **Pclass**.

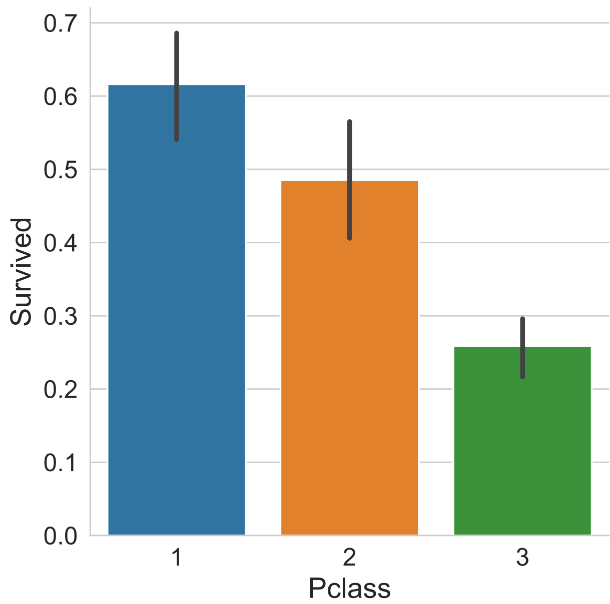

Figure 5.2: Proportion of survived passengers in each **Pclass**

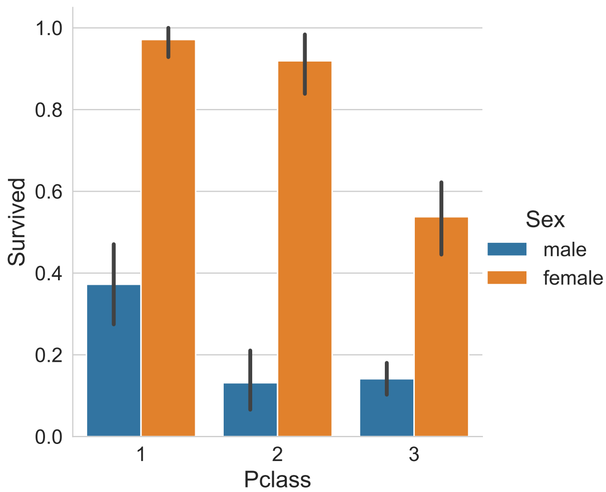

Figure 5.3: Proportion of survived passengers of eache gender in each **Pclass**

5.2 Name

The **Name** attribute contains names of passengers, and it contains 712 unique values in the exploratory dataset. The table 5.1 shows the exerpt from this column. There seems to be a useful pattern.

Each name contains a title, such as "Mrs." or "Master.". The title may contain information about gender, socio-economic status, age, etc. Let's extract the title from each name and count how many times each title occurs in the exploratory set. The results is shown in the table 5.2 and figure 5.4.

Several titles occurs extremely rarely, later we will combine them with more common titles.

Table 5.1: Excerpt from the **Name** column of the exploratory set

	Name
788	Dean, Master. Bertram Vere
347	Davison, Mrs. Thomas Henry (Mary E Finck)
629	O'Connell, Mr. Patrick D
734	Troupiansky, Mr. Moses Aaron
106	Salkjelsvik, Miss. Anna Kristine

Table 5.2: Number of occurrences of each title

Title	**Number**	**Title**	**Number**
mr	415	jonkheer	1
miss	144	ms	1
mrs	88	the countess	1
master	29	don	1
dr	6	mme	1
rev	4	sir	1
major	2	capt	1
col	2	mlle	1

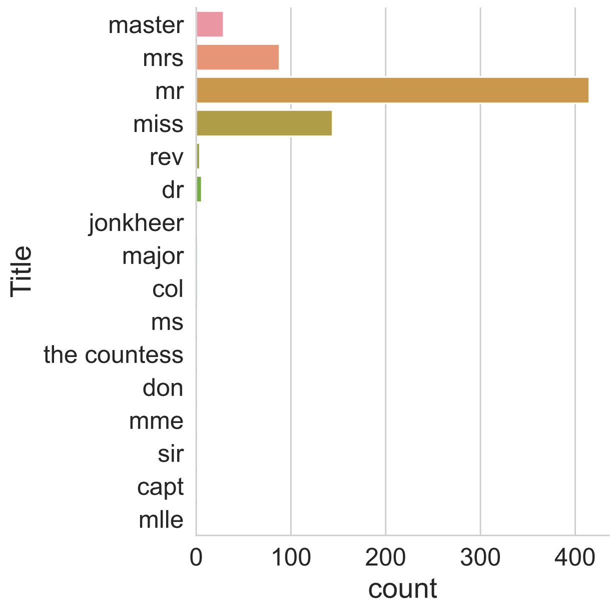

Figure 5.4: Number of occurrences of each title

5.3 Sex

Exploratory Analysis

5.4 Age

Exploratory Analysis

5.5 SibSp

Exploratory Analysis

5.6 Parch

Exploratory Analysis

5.7 Ticket

Exploratory Analysis

5.8 Fare

Exploratory Analysis

5.9 Cabin

Exploratory Analysis

5.10 Embarked

Exploratory Analysis

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).