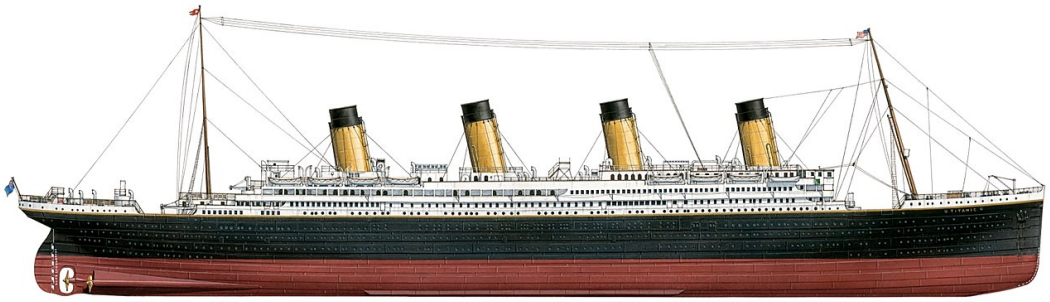
Titanic: Machine Learning from disaster



Emanuele Fittipaldi

2021

Contents

1 Data Set description

This is a famous data set in the kaggle community. This data set is involved in the Titanic ML Competition which is described as the best first challange for a beginner to dive into the ML. Even though the aim of this competition is predicting which passengers survived the Titanic shipwreck, for the purpose of this Course I am going just to do all the steps in the machine learning workflow that preceeds the actual training of the model.

1.1 Data Dictionary

- **survival** - If the person survived or not |0 = No, 1 = Yes.
- **pclass** - Ticket class
- **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

Types of the variables.

- Survived - int64
- Pclass - int64
- Name - object
- Sex - object
- Age - float64
- SibSp - int64
- Parch - int64

- Ticket - object
- Fare - float64
- Cabin - object
- Embarked - object

in this data set there are 891 observations ditribuited across 12 columns. A possibile target variable is 'Survived'. Tha analysis i am going to do is based on that.

1.1.1 Variable Notes

- **pclass** - A proxy for socio-economic status
- **age** - Age is fractional if less than 1. If the age is estimated, is it in the form of xx.5
- **sibsp** - The dataset defines family relations in this way... Sibling = brother, sister, stepbrother, stepsister Spouse = husband, wife (mistresses and fiancés were ignored)
- **parch** - The dataset defines family relations in this way... Parent = mother, father Child = daughter, son, stepdaughter, stepson Some children travelled only with a nanny, therefore parch=0 for them.

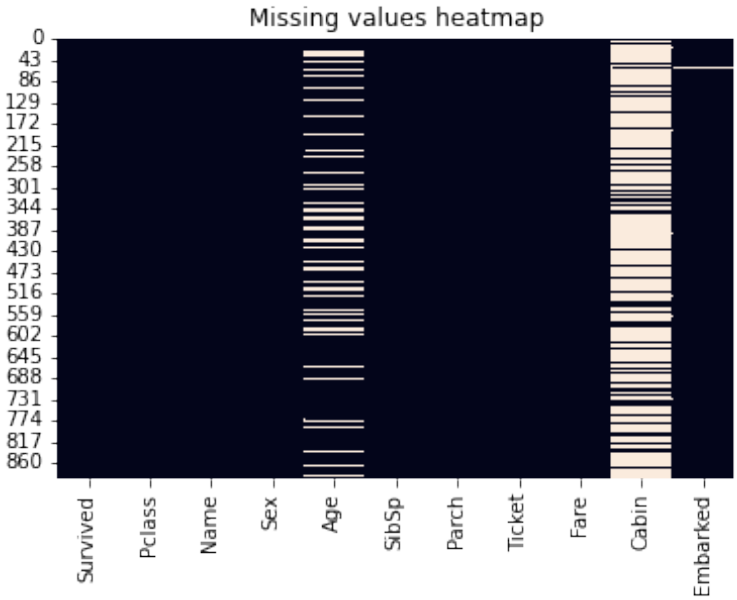
2 Data Set Exploration

2.1 Data Set Exploration Plan

I am going to esplore the Data Set in two ways:

- **Statistically** - I am going to find all the useful stats about the data set as
  - Column names, Column types, # observations, # columns, dtype of the columns
  - Mean, Median, Interquartile ranges, Min/Max range, standard deviation
  - Correlation Matrix
- **Visually** - Printing the data visually makes it clearer the understanding of patterns and correlations that might exist in the data.
  - Histograms
  - Scatter Plots
  - Pie Charts
  - Box Plots
  - Bar Plots

2.2 Summary of the data



Observation: Thare are alot of null values in the Cabin Column, and a considerable amount in Age Column. These are things to be accounted in the Data Transfrmation stage. Embarked has only a missing value, I could drop the row since it is a single observation or preserve it by filling the missing value. *Note that at this point I've already dropped the column PassengerId since It was only and index, carrying no useful information.*

	count	mean	std	min	25%	median	75%	max	range
Survived	891.0	0.383838	0.486592	0.00	0.0000	0.0000	1.0	1.0000	1.0000
Pclass	891.0	2.308642	0.836071	1.00	2.0000	3.0000	3.0	3.0000	2.0000
Age	714.0	29.699118	14.526497	0.42	20.1250	28.0000	38.0	80.0000	79.5800
SibSp	891.0	0.523008	1.102743	0.00	0.0000	0.0000	1.0	8.0000	8.0000
Parch	891.0	0.381594	0.806057	0.00	0.0000	0.0000	0.0	6.0000	6.0000
Fare	891.0	32.204208	49.693429	0.00	7.9104	14.4542	31.0	512.3292	512.3292

Observation: According to train.csv, the youngest onboard was a 5 months baby. The oldest 80 years old. The first thing I noticed in these stats is the max value for the Fare column. It was very different compared to the average. In the beginning I thought that there must be an outlier throwing my mean off, but after I observed that it was a legit Fare. Those cabins, B51,53,55 were one of the most expensive. Since It affect only 3 observations, I think I could remove these values afterwards.

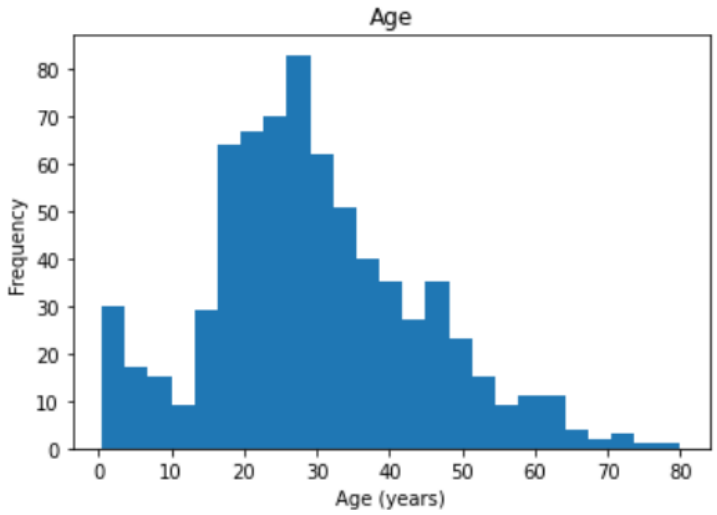
	Survived		Pclass		Age		SibSp		Parch		Fare	
	mean	median	mean	median	mean	median	mean	median	mean	median	mean	median
Sex												
female	0.742038	1	2.159236	2	27.915709	27.0	0.694268	0	0.649682	0	44.479818	23.0
male	0.188908	0	2.389948	3	30.726645	29.0	0.429809	0	0.235702	0	25.523893	10.5

Observation: the arithmetic mean is the only value that can be mislead by the presence of an outlier. Conversely the median is not influenced. I am reporting those values for each column to see if there are huge differences between them. If that's the case, this could indicate that there is an outlier, otherwise if the median and the arithmetic mean are close, there shouldn't be any. In fact if we observe the difference between mean and median for the Fare column we can see that the mean is considerably larger. This is due to the expensive tickets discssed above that lead the mean to increase.

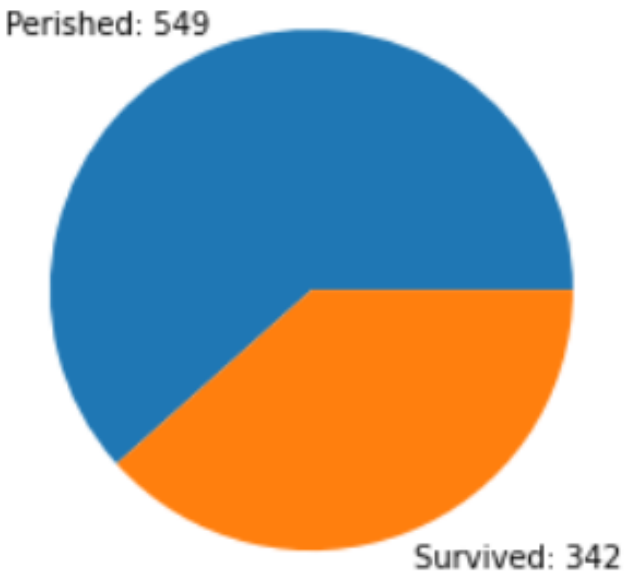
Observaton: Different family members could travel under the same ticket number.Ordering the rows by ticket number I gave a look to some of these families since the members of the same family are now consecutive in the DataFrame.

Observaton: At first I tought that there was an outlier in the Fare column. Afterwards i discovered the story behind these expensive tickets. They belonged to Thomas Drake Martinez Cardeza a very wealthy banker who travelled with his wife, mother and his manservant. They all survived, even though the manservant committed suiced few years after the shipwreck.

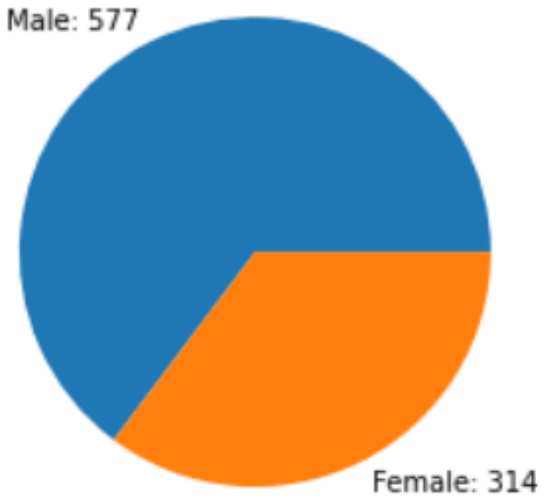
Observaton: There are free tickets also for third class members because they were American Lines employees (LINE) and they've been offered the ticket for free.



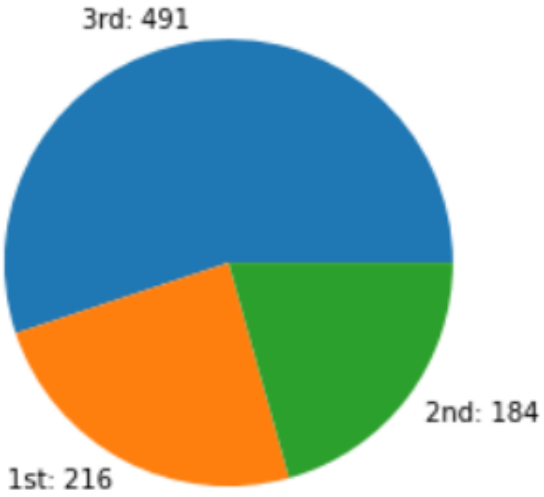
Observaton: The majority of people on the Titanic was under 30.



Observation: The 61% of people died. The 39% survived.



Observation:The 64% of the passengers were male vs. 36% female

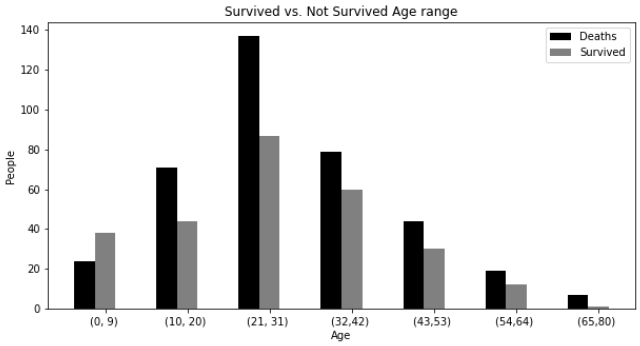


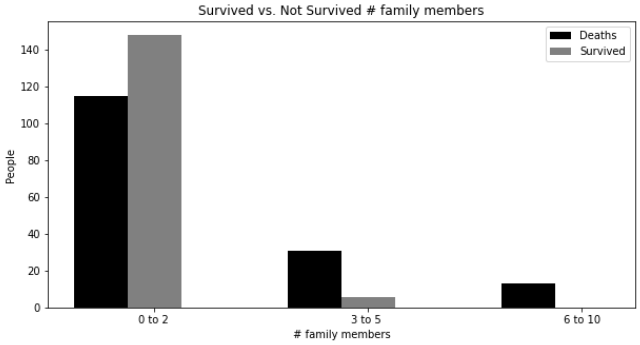
Observation:

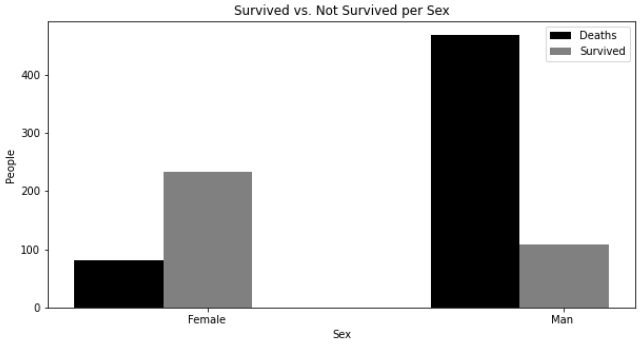
- 55% Third class.
- 21% Second Class.
- 25% First class.

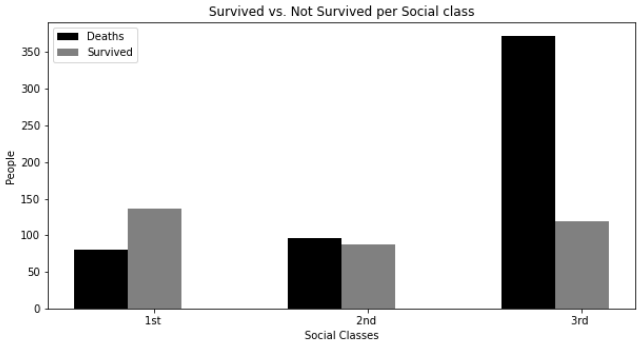
3 Data Cleaning and Feature Engineering

At this point I did very little or no data cleaning. In this section I am going to tackle missing values, feature encoding, feature engineering in order to find some useful correlations that could carry a good predicition power.









Observation: Since the Columns Sibsp & Parch referres basically the number of relatives of a person, I created a new column called Family which indicates how many people travelled with that person. I decided to that in advance in order to explore the data set even further, according to this new information.

Observation: The females survived the most. Is It because of "first womans and childs" thought?

Observation: We can see that in the range 0-9 years of Age, this is the only range where people survived the most. So it is plausible to think that the rule "babies and woman first" was really observed.

Observation: This shows that if you travelled with lass than 3 family members your chances of survival increased exponentially.

Obeservation:

- 1st class members are the only one where the number of people who survived are more than the dead ones. Is It because of a "rich folks first" thought? or because the icebarg struck where the 3rd class cabins were located?

- 3rd class members had the worst.

I added a new column Is_Alone in order to see if there is a strong correlation between being alone and the chances of surviving.

3.1 Data Imputation

Observation: Since 'Embarked' only had two missing values and the largest number of commuters embarked from Southampton, the probability of boarding from Southampton is higher. So, we fill the missing values with Southampton.

Observation: I tackled the filling of the null values for the Age column. I've used the mode to fill these missing values. The mode is the most frequent value in a list of values. Afterwards I added a Youth feature which indicates the range of age in which the person fall. I did this since It was observed that people between 21 and 31 years old where the ones who perished the most. The renges of youth are:

- 0 to 9 -> Child
- 10 to 20 -> Teenager
- 21 to 31 -> Adult
- 32 to 55 -> Senior
- 56 to 80 -> Old

After that I did drop the columns I wouldn't need anymoe (Ticket, Name, Cabin).

3.2 Encoding

At this point I had to encode the categorical features. I use LabelEncoder from Sklearn to encode the Sex and Is_Alone lables automatically. LabelEncoder gives a unique number for every label values it finds in the given column. On the other hand I've used one-hot encoding to encode 'Youth','Embrked'.

4 Hypotesis formulation

- Ho: μ Age_Survived == μ Age_Not_survived
- Ha: μ Age_Survived != μ Age_Not_Survived
- Ho: μ Fare_Survived == μ Fare_Not_survived
- Ha: μ Fare_Survived != μ Fare_Not_Survived
- Ho: μ Class_Survived == μ Class_Not_survived
- Ha: μ Class_Survived != μ Class_Not_Survived

5 Significance test

The Kruskal-Wallis H-test tests the null hypothesis that the population median of all of the groups are equal. It is a non-parametric version of ANOVA. The test works on 2 or more independent samples, which may have different sizes. Note that rejecting the null hypothesis does not indicate which of the groups differs. Post hoc comparisons between groups are required to determine which groups are different.

- H-statistic: 0.47409198473728537
- P-Value: 0.4911106630298028
- Accept NULL hypothesis - No significant difference between groups.

This very interesting. According to the Kruskal-Wallis test there is no difference between the age of the people who survived and the age of the people who did not survive. So we can conclude that these samples belong to the same probability distribuition.

- H-statistic: 92.86163001934372
- P-Value: 5.608129813369624e-22
- Reject NULL hypothesis - Significant differences exist between groups.

We rejected the null hypotesis. So there Fare had certainly an impact on the people who survived and the ones who perished.

- H-statistic: 102.6831333974126
- P-Value: 3.932785644652686e-24
- Reject NULL hypothesis - Significant differences exist between groups.

As we expected, also the Class had an impact on the survival rate The reject of the null hypotesis is the proof.

6 Suggestions

I could suggest these steps in order to further analyzie the data.

- before and after plots acccording to the transformations I did so far.
- Analyze the data set considering male vs female and not survived vs perished.

7 Sumary

Overall this data set was not that bad. The biggest problem was just the presence of null values, which some required a little bit more work than others. Another added difficulty I could mention about this data set, was the necessity to deal with many categorical data that required properly encoding and the fact that the target variable was basically a boolean (0 or 1). This made the scatter plots not very useful and the finding of special patterns/trends very difficult.