Kaggle - Titanic Dataset

Alfonso Ponce Navarro

August 2022

Contents

1 Dataset Description **1**

2 First Version **2**
2.1 Procedure 2
2.1.1 Attribute elimination 2
2.1.2 Missing Values detection 2
2.2 Result 6

3 Second Version **7**
3.1 Correlation analysis 8
3.2 Result 8

4 Correction about "validation" accuracy **8**

5 Third Version **9**
5.1 Reconsideration of imputation method 9

6 Third and last version **10**

1 Dataset Description

The Dataset contains 12 attributes with multiple instances. Attributes and their descriptions can be found in `https://www.kaggle.com/competitions/titanic/data?select=train.csv`
It is a classification problem in which we are asked to predict if a passenger survived or not during titanic sink.

2 First Version

2.1 Procedure

2.1.1 Attribute elimination

First of all, after reading the attributes descriptions, I considered to remove some attributes that will not add any relevant information, below it is justified their removals.

2.1.1.1 Passengers Name

Apparently the passengers name does not tell us anything useful to predict if he/she survived or not. But, if we investigate, maybe the name could tell us if the passenger was rich or not and thus we could predict the class of ticket he/she bought.
But also in the dataset we are given the class of the passenger, so his/her name is just irrelevant.

2.1.1.2 Ticket

As mentioned above, we are given the class of the passenger, so a priori this attribute is also irrelevant.

2.1.1.3 Passenger ID

This is just an index, not useful.

2.1.1.4 Cabin

The Cabin number may be relevant, but for the first version, we will base our prediction related to the position of the passengers in the ship on the ticket class.

2.1.1.5 Embarked

The port of Embarkation can not be the cause of the survival of a passenger. Maybe, we could find a correlation but that does not imply causality.

2.1.2 Missing Values detection

The second intuitive idea was to find if in the resultant dataset was any missing value. After a search, the unique attribute which had any missing value was the Age of the passengers.
There where 20% of rows with null Age. That is not a low percentage, so erasing incomplete rows is a very bad idea. Firstly, I explored if missigness is biased.

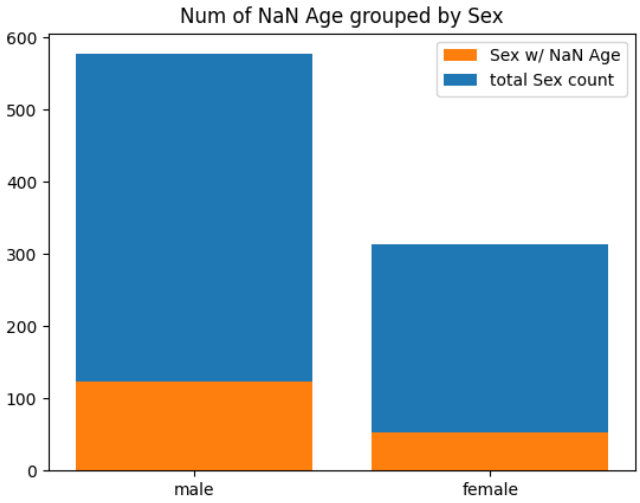

Figure 1:

As we can see in the plot more men have not their age been recorded. In relative terms, we can see that:

- 21% of male passengers have their age missingly recorded.
- 16 % of female passengers have their age missingly recorded.

This is a little bit strange since culturally women are more likely to refuse to tell their ages. Let's see if there has been another cause of this lack of information.

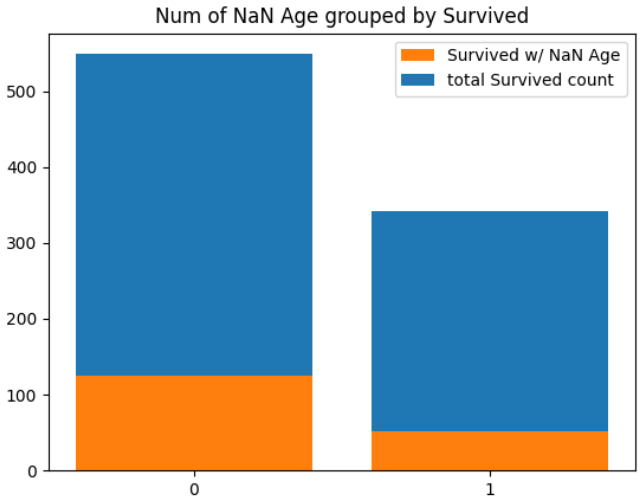

Figure 2:

Now I wanted to test if the age was not recorded due to survival of the passengers. In the X axis *0* indicates the dead passengers and *1* indicates those who survived. Relatively speaking:

- We don't know the age of the 22% of dead people.
- From the survivors, we lack 15% of their ages.

It is greater the amount of people we don't know their age in the dead group than in the survived one. This could make us think that age was recorded after the sink, when recoveries of the corpses were done. The lack of information from living people might be because registers were lost during time.
But also I have take into account another variable, the ticket class. This can be one of three (1st class, 2nd class or 3rd class). Now this graph shows more interesting information:

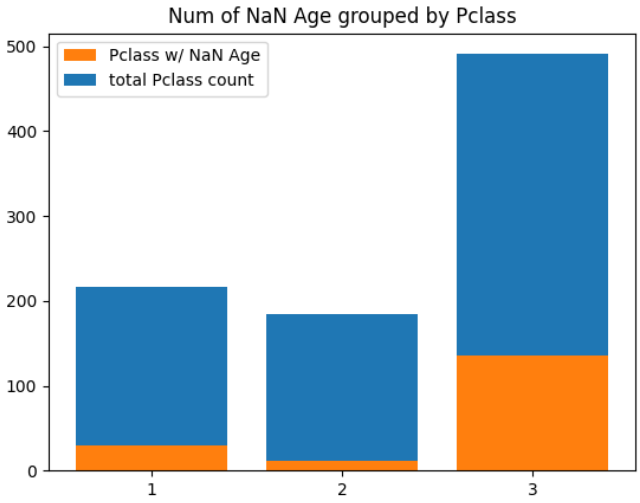

Figure 3:

As I have been doing, I will show in relative terms that:

- 13% of 1st class ticket have their age unknown
- 5% of 2nd class ticket have their age unknown
- 27% of 3rd class ticket have their age unknown

It is very significant the difference between the third class (the worst one) with the other in terms of age missingness. Lets see now the deaths related to the ticket class.

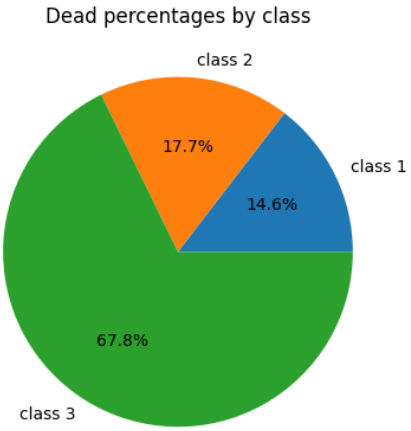

Figure 4:

In effect, most part of dead people were in the third class, the same class that has most people with unknown age.
Taking into account that there are not significant differences between the gender age missingnes, but knowing that the class with the greatest death rate is the same as the class with most of missingness recorded age, it is not crazy to affirm that age was recorded after the sink, probably when recovering the corpses.
Data has sense if we think that probably lifeboats were limited and reserved mostly for high class passengers.

After all of these we can now think about the algorithm to impute the missing values.
To select the imputation algorithm it is important to take into account the classification algorithm to use because some imputation methods work better with one classifier than with another. We can go for the simplest approach since we dont require computation speed, just good accuracy and we have 10 attempts, so KNN is not a bad idea for classification.
For this classifier, a good imputation approach is **Conditional Mean Completion**, that is, imputing values through computing the mean of specified class.

2.2 Result

After doing all the procedure, the accuracy obtained is 0.66267.
This is not a good result but still improvable. Now, our aim is to improve that result, that is not difficult since we have a lot of margin to maneuvre with more

sophisticated procedures.

3 Second Version

Now, we are going to see the difference of performance of our model between the train data that we are provided and the test data. From all the training data, we are going to (pseudo)randomly sample a 30% for "validation" (it is not exactly validation since we are not going to select a best model).
For doing this experiment, we maintain our KNN model configuration (K=10) and we will repeat all the procedure for 500 times. This is because statistically, it can provide significant results.
We reach a 75% of mean accuracy with a standard deviation of 2%. In the below QQ-plot we can see that the sample of accuracies fit approximately a normal distribution, so our mean should be a good representative of our sample.

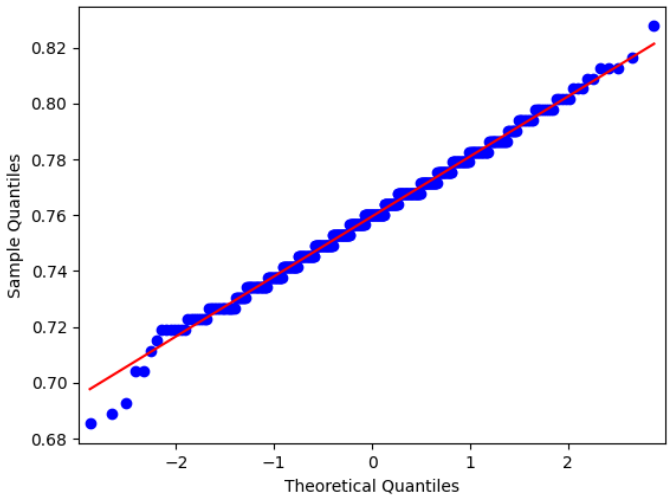

Figure 5: Caption

Moreover, after applying the Shappiro-Wilk normality test, we can assure that our sample of accuracies with a confidence of 99.5% follows a normal distribution. So, if statistically our estimation is confidently in a 76%, why in the competition the model scored a poorly 66%?
If we analyze the test dataset, we can see there is two columns with null values instead of one: Gender (again) and ticket's fare. Since we impute null values with conditioned mean, this may alterate the generability of our model, for example, alterating the underlying distribution of the ticket's fare.
Another question arises, is fare a valuable variable for predicting the survival of

passengers or it is just redundant, or moreover noise? Can we discard it? Now it is time to find if there is a pattern between fare and survivality, and if we find it, if it the same of, for example, the ticket class.

3.1 Correlation analysis

Intuitively, we can think that if there is any correlation between the fare and any other variables, the most likely one to be correlated might be the ticket class since it is a good indicator of ticket price(it is not a universal rule, i.e., there are free 1st class tickets).
But this is not a typical correlation in which we can apply pearson correlation or even spearman since we have one continuous variable(Fare) and one categorical variable(ticket class). So, there cannot be any linear or non-linear relationship between them.
One simple way to achieve this task is to analyze the variance of the overall continuous variable and then compare it into the variance after grouping the continuous variable according to the categorical label. If the difference is very significant, it may mean that the categorical data can explain most of the continuous variable and thus indicate a strong correlation.
An overall variance of 2466.66 is obtained.
Then the variance of 3rd class, 2nd class and 1st class are 138.44, 179.04 and 6115.04 respectively.
As we said before, we can think that the ticket classes can really explain the fare of tickets as they differ very much from the overall. This is not a very strong statistical proof but it may be sufficient to state that fare and ticket class are very associated. So, having this statistical background, the decision taken is to erase the fare variable.

3.2 Result

It is obtained a 0.70. It is not an incredible score but may we have done a good approach. Also we have not explored any other algorithms neither even done a validation step. But there is still a remarkable thing.
If we compute the "validation" step, we encounter that we score a 0.84. There is a very high improvement, but still far from the data test.

4 Correction about "validation" accuracy

After doing some research, test set and train set are confirmed both to have the same distributions, so it cannot be that what it is making such difference in accuracy's values.
So what I did was to investigate the code for the existence of some bugs. And that was it the key problem. When random sampling was done to check for accuracy, there was a big mistake, that is, not erasing those instances that were in validation set, in the training set, so obviously, the validation score would be

high. When the bug was fixed, approximately a 73% percent was reached. Theorically, it is known that simple random sampling have low variability due to repetition of instances. So to have more realistic results, a K fold cross-validation will be performed.

5 Third Version

5.1 Reconsideration of imputation method

At first glance, we used the mean for age imputation due to the use of a distance based classification as KNN. Now, we should think a little bit further, is it really correct to use the mean blindly? Actually it is not. We must see first the distribution of the ages because if we have an skewed distribution, its mean will not be significant and thus be not useful for calculating the ages. Lets see the distribution of ages in our training set.

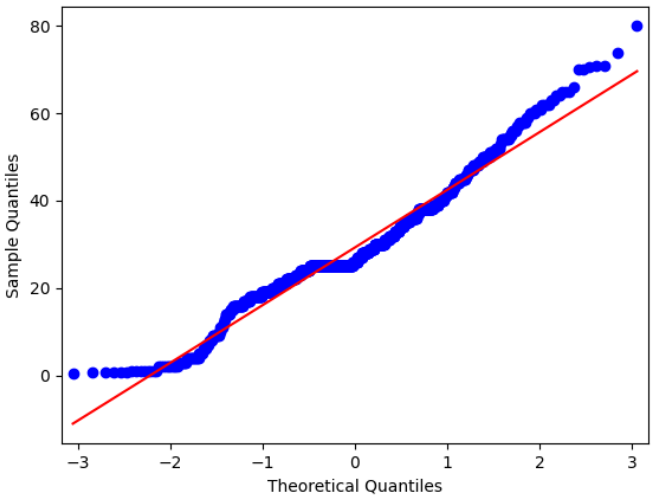

Figure 6: Caption

As we can see in the QQ plot, there is a departure in ages, specially significant as ages decreases. If we perform a Box Cox transformation, that is, transforming a non normal distribution to normal one, we have the following QQ plot:

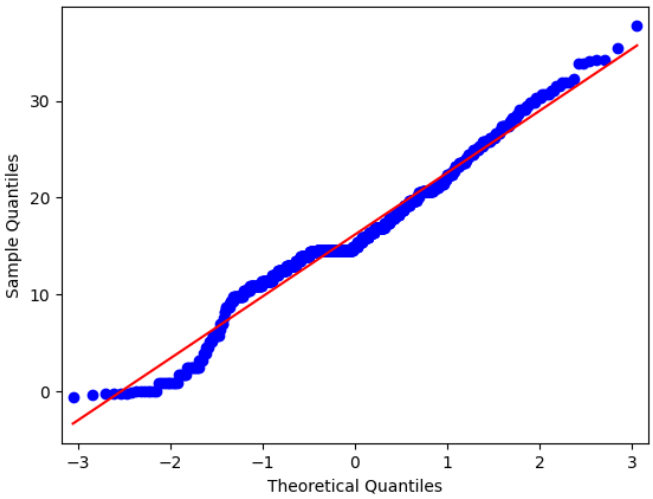

Figure 7: Caption

Now our data is better fitted to a normal distribution, but the departures in lower ages still exist. If we use the Shapiro normality test we obtain that our data remains not gaussian, so the Box Cox transformation has not done great impact. We now encounter two branches:

- Find another imputation method since the mean of our distribution is not reliable.
- Detect noise and deal with it.

After doing some research on Internet, people mainly do CMC imputation, so I will remain that method.

6 Third and last version

Since I don't want to dedicate so much time to this dataset, I decided to go to the final part, select the best algorithm to achieve the best accuracy.
The following algotithms were tested:

- KNN – 72%
- Decision tree – 74%
- Kernel SVM (Since class are not linearly separable, the use of a kernel is a must) – 76%
- XGBoost 77%

Hyperparameter Tunning was done using Grid Search for obtaining the best hyperparameter combination from all the possibilities I stablished.

Although we achieve the best accuracy with XGBoost I did not stop trying some alternatives.
I wanted to build a simple multiclassifier system based on the four algorithms described above. The final and best result obtained is a 78%.