**STORYTELLING - TITANIC, MACHINE LEARNING FROM DISASTER**

A century has sailed by since the luxury steamship RMS Titanic met its catastrophic end in the North Atlantic, plunging two miles to the ocean floor after sideswiping an iceberg during its maiden voyage. Rather than the intended Port of New York, a deep-sea grave became the pride of the White Star Line’s final destination in the early hours of April 15, 1912. More than 1,500 people lost their lives in the disaster. In the decades since her demise, Titanic has inspired countless books and several notable films while continuing to make headlines, particularly since the 1985 discovery of her resting place off the coast of Newfoundland. Meanwhile, her story has entered the public consciousness as a powerful cautionary tale about the perils of human hubris. And what we know about this giant ship, Titanic had a length of 882 feet long and a weight of 46.000 tons, 9 decks and height about 175 feet. On board 2.224 passengers and crew (3.500), but lifeboats for only 1.178 people, incredible! 20 boilers & 162 furnaces burned through 650 tons of coal per day, cranking out 16.000 horsepower. Top speed 24 knots, 4 smokestacks each 22 feet wide and 6 stories high. It took 3.000 people over 3 years to build Titanic. In 1912 first class ticket cost 4300 dollars and third class ticket 36 dollars. Titanic was cruising at 22 knots when it hit an iceberg on april 14, 1912 at 11:40 PM.The iceberg cut a hole into the hull between 220 and 250, finally, Titanic sank in less than 3 hours.The survival rate: first class 60%, second class 42%, third class 25% and Crew less than 25%. Titanic’s final resting place 350 miles southeast of newfoundland, Canada, 2.3 miles below the ocean’s surface.Turning back to reality and our dataset, we have Data for Training and Testing, we have a description about the variables:

VARIABLE DESCRIPTIONS:

survival Survival

(0 = No; 1 = Yes)

pclass Passenger Class

(1 = 1st; 2 = 2nd; 3 = 3rd)

name Name

sex Sex

age Age

sibsp Number of Siblings/Spouses Aboard

parch Number of Parents/Children Aboard

ticket Ticket Number

fare Passenger Fare

cabin Cabin

embarked Port of Embarkation

(C = Cherbourg; Q = Queenstown; S = Southampton)

Now take a look and listen what the Data tell us, first training data set:

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 891 entries, 0 to 890

Data columns (total 12 columns):

PassengerId 891 non-null int64

Survived 891 non-null int64

Pclass 891 non-null int64

Name 891 non-null object

Sex 891 non-null object

Age 714 non-null float64

SibSp 891 non-null int64

Parch 891 non-null int64

Ticket 891 non-null object

Fare 891 non-null float64

Cabin 204 non-null object

Embarked 889 non-null object

dtypes: float64(2), int64(5), object(5)

memory usage: 83.6+ KB

On checking for missing data, we found:

Total %

Cabin 687 77.1

Age 177 19.9

Embarked 2 0.2

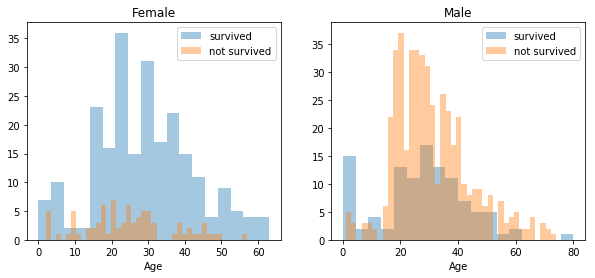
Fare 0 0.0

Ticket 0 0.0

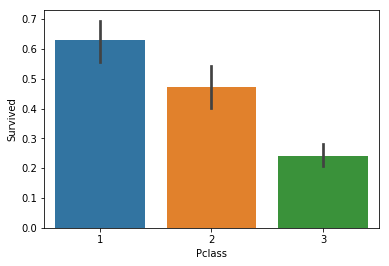
In this case it’s important that you fill the NA’s values to have a better fit in your model, you can do it with several methods, for example: Replacing the missing values with the average of the available values. Or in some cases, just dropping the missing values.

Next, we could find correlation of different variables with survival.For eg.:

**Age and Sex**



**PClass**



A 41% of third class died, third class was the 55% of the passangers, and 15% of first class survived.Well you can see inside the Data a lot of information and if you try to understand the situation you will imagine yourself living the tragedy and trying to understand what happened.

Next, we check for datatypes before applyimg ML models.

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 891 entries, 0 to 890

Data columns (total 11 columns):

Survived 891 non-null int64

Pclass 891 non-null int64

Name 891 non-null object

Sex 891 non-null object

Age 891 non-null int32

SibSp 891 non-null int64

Parch 891 non-null int64

Ticket 891 non-null object

Fare 891 non-null float64

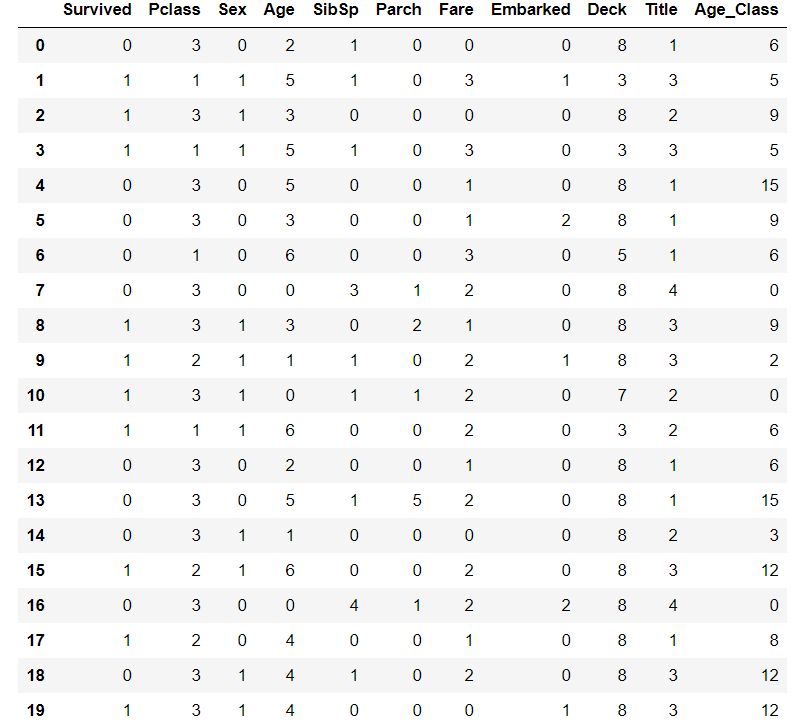
Embarked 891 non-null object

Deck 891 non-null int32

dtypes: float64(1), int32(2), int64(4), object(4)

memory usage: 69.7+ KB

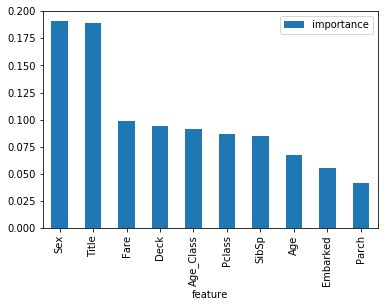
As some datatypes are different so converting them into required datatypes.Finally, checking our training set before evaluation.



Next, we apply different ML models on the dataset. Comparing them we find:

|  | **Model** |
| --- | --- |
| **Score** |  |
| **92.70** | Random Forest |
| **92.70** | Decision Tree |
| **80.81** | Logistic Regression |
| **80.58** | Support Vector Machines |

Applying K-Fold cross validation we find our model has a average accuracy of 81.7% with a standard deviation of 3.72 %. The standard deviation shows us, how precise the estimates are.This means in our case that the accuracy of our model can differ + - 3.72%. Using sklearn, we can find the feature uimportance of each of the corresponding features.



Now, on observation we could remove Parch due to its low effect on the data. If we train the classifier again our Random Forest gives an accuracy of 92.7%. Most of the hyper parameters regulate the construction of the conditional inference trees. So,we can tune these features using Grid Search.

Still, there is a lot of opportunity for improvement in this implementation.At this moment **we can observe the potential of Data Science but this is only the top of the iceberg**.

