Titanic: Machine Learning from Disaster

Introduction
The sinking of the Titanic is one of the most infamous shipwrecks in history.

On April 15, 1912, during her maiden voyage, the widely considered “unsinkable” RMS Titanic sank after colliding with an iceberg. Unfortunately, there weren’t enough lifeboats for everyone onboard, resulting in the death of 1502 out of 2224 passengers and crew.

While there was some element of luck involved in surviving, it seems some groups of people were more likely to survive than others.

In this challenge, we were asked to build a predictive model that answers the question: “what sorts of people were more likely to survive?” using passenger data (i.e name, age, gender, socio-economic class, etc).

Objective
The goal is to build a Model that can predict the survival or the death of a given passenger based on a set of variables describing their attributes such as age, sex, or passenger class on the boat.

Methodology
The type of machine learning method used is called classification, because predictions were classified among each passenger as ‘survived’ or not.

Two datasets were extracted from Kaggle **- train set** and **test set**, which have a data set of different information about passengers onboard the Titanic, and we used that information to predict whether those people survived or not.

The **training set** contains data we used to train our model. It has a number of feature columns which contain various descriptive data, as well as a column of the target values used to predict: in this case, Survival. The **testing set** contains all of the same feature columns, but is missing the target value column. Exploring through the data indicates that Age, Sex, Pclass, Parch and Sibsp would be a good predictor of survival.

Logistic Regression was used to train our model and scikit-learn library tools was used to instantiate, fit, predict and evaluate the accuracy of the predictions .

To give us a better understanding of the real performance of our model, a technique called **cross validation** was used to train and test our model on different splits of our data, and then average the accuracy scores.

Result
Analysis and visualization of data can be seen on titanic.ipynb file attached to the link of this file.

Conclusion
From the results of our k-fold validation, the accuracy number varies with each fold – ranging between 76.4% and 85.3%. As it happens, our average accuracy score was 80.5%, which is not far from the 80.4% we got from our simple train/test split.

At the time of submission on kaggle, my prediction received a ranking of 79.9%