1. Introduction

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, I am going to build a predictive model that answers the question: “what sorts of people were more likely to survive?”

1. Data Analysis

Feature Overview:

survival - if the passenger survived, 0 is No, and 1 is Yes

Pclass - ticket class, 1 is 1st, 2 is 2nd and 3 is 3rd

Sex - gender sex

Age - age in years

Sibsp - number of siblings / spouses

Parch - number of parents / children

Ticket - ticket number

Fare - passenger fare

Cabin - cabin number

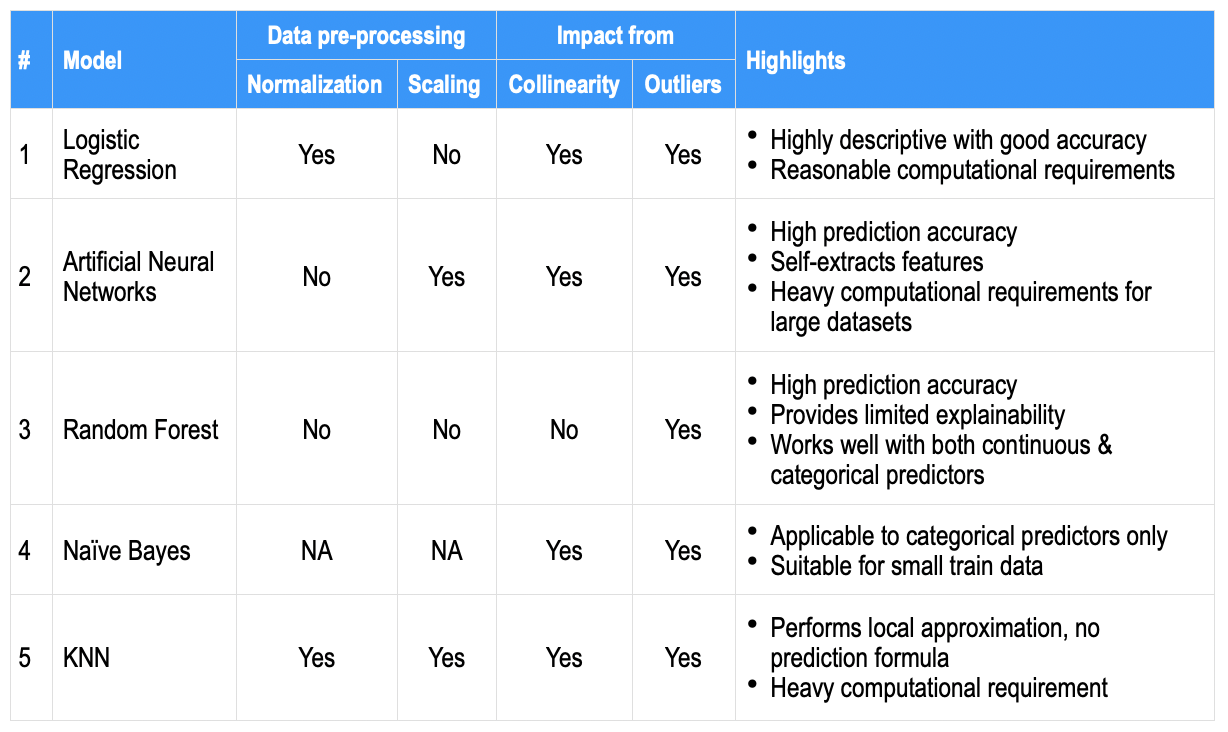
Embarked - port of embarkation, C is Cherbourg, Q is Queenstown, and S is Southampton

Note: pclass is associated with socio-economic status, age is fractional if less than 1 and if is estimated, it is in the form of xx.5.

1. Problem Executive

The aim is to create a supervised classification model for predicting if the passenger will survive, the data feature is relatively simple thus consider normally used machine learning model, such as random forest, logistic regression, K nearest neighbour, and Naive bayes.

Trade-off analysis:



Note: Artificial Neural Networks on the second row should be deep neural networks.

Ref: <https://www.analyticsvidhya.com/blog/2020/11/popular-classification-models-for-machine-learning/>

As the data contains both numerical and categorical features, the random forest is the main model. However the data is relatively small, the decision is to try all of the classification model and make a detailed comparison.