EN502 - Machine Learning
Initial Project Report

"Titanic - Machine Learning from Disaster"

Tanuj Raghav	**Pragyan Jaiminy**	**Pushpak Prateek**
19-11-EC-027	19-11-EC-028	19-11-EC-039
tanuj81_soe@jnu.ac.in	pragya32_soe@jnu.ac.in	pushpa89_soe@jnu.ac.in

---

One of the most ill-famed incidents in the history of shipping is the sinking of the Titanic. It sank after colliding with an iceberg. This resulted in the unfortunate death of 1502 out of 2224 passengers, including crew members, mainly because of the unavailability of the lifeboats. Some were lucky to survive. Further research and investigations showed that luck was not the only factor that decided the fate of the travellers. The socio-economic class, gender, age, family size, lodging, etc. too affected the outcome. This project aims to explore and analyse the effect of these factors on the chance of survival of a person on board. Then we predict if a passenger survived or not using the model thus trained.

MOTIVATION

We based the project on the infamous past scenario. External factors were unavoidable, but several other factors which affected the survival probability could have been controlled. This made us inclined to work on finding the extent of the impact of the mishap.

CHALLENGES

Our first challenge was the data provided, as it wasn't very machine friendly. The next difficulty was selecting the best algorithm as per our requirement of achieving higher precision. Also, there were many features which required proper usage of feature engineering. Further, the model had to be trained and tested several times to reduce the percentage error. Also, we had to reduce model-based errors such as false positivity, false negativity.

METHODOLOGY

The very first step in our pipeline would be to get the data. We would ask ourselves various questions before collecting data to get a smaller and precise set of data. The second step will prepare the data. This step would involve cleaning the data. Empty or missing data entries would either be filled or removed. The next and third step involves data pre-processing. In this step, data is transformed, encoded, and made in such a form by which the machine can easily pursue it. The algorithm will easily interpret the features of the data. We further move on to analyse the passenger manifest. This will give us more of an idea about which direction we should head. Once we are done with preparing and pre-processing. We will now split the data

into three parts for three different usages. One will be to train the model, second will be to validate and the third one will test the accuracy of the model. Then we train the model. We will apply various algorithms and note their accuracy of prediction. We chose the one with the highest accuracy.

RESULTS

We expect to chart out the dependence of the survival rate on the factors mentioned in the problem set. We expect to conclude that people who travelled in the first class should have a higher rate of survival as compared to the other two classes. We expect the embankment port doesn't matter much with survival rate. Also, we predict that age and qualification would be a prominent factor for better chances of survival.

The model made will be judged based on various performance metrics, such as sensitivity, precision and accuracy. Precision is defined as the number of correct documents returned by the model made. Sensitivity, which is also called as Recall, is the number of positive returns by our model. Also, we would compare our works with other submissions on Kaggle and check where we stand in real life.