Titanic: Distributed Machine Learning from disaster

GIOVANNI MANFREDI SEBASTIANO MENEGHIN
gioman | meneghin @kth.se

30th September 2023

1 Problem

Last summer the tragic implosion of the Titan submarine operated by the company OceanGate, revamped the attention over disaster happening also to billionaires. Even if very rare, these events can happen and predicting them would be beneficial to many human lives.

The goal of this project is to process and query data regarding the historical Titanic wreckage [1] and implement a predictive analysis to classify which of the passenger on board will survive the tragedy, based on various parameters (name, age, gender, social status, etc.). The structure in Spark for pre-processing and querying for the model training will be the main output of this project.

2 Tools

The project will make use of:

- Spark and Spark SQL to pre-process and query the data.
- Spark ML for the prediction model's implementation.

3 Data

The project will use the data set provided by the Titanic - Machine Learning from Disaster [1] online competition.

4 Methodology and algorithm

The project will first download the data from the Kaggle website, then using the Spark environment pre-process it to improve the quality of the data set. The data set created in this way will be queried to ensure the correct behaviour of the Spark environment.

Then, after first dividing the data set into training, validation and test set, multiple classification algorithm will be trained. To select the best technique the models will be tested on the validation set. Finally, the results of the selected model will be tested using the test set.

References

[1] W. Curkierski, "Titanic - Machine Learning from Disaster — kaggle.com," https://www.kaggle.com/c/titanic, 2012.