Titanic Kaggle Challenge: Addressing Data Processing for Improved Survivability Prediction

Toan Thang Nguyen, Tung Thien Duong, Minh Tien Ho, Thai Tu Dang

Saigon University, Vietnam.

**Abstract**

Titanic challenge, hosted by Kaggle, is a common challenge that is suitable for beginners who are not familiar with training models. However, it is not an easy task to create a model that could predict high scores since the provided dataset has lots of missing values and seemingly unnecessary features, which could make it harder for the model to be trained efficiently. This study employs data processing techniques to address these issues, to facilitate more effective model training. [Sẽ cập nhật các kỹ thuật được sử dụng trong bài báo này sau.]

**Keywords:** survivability prediction, data processing

# 1 Introduction

The sinking of the RMS Titanic on April 15, 1912, remains one of the most tragic maritime disasters in history, resulting in the loss of 1,502 out of 2,224 passengers and crew. In the aftermath, researchers have sought to understand survival factors, leading to the widely recognized Titanic challenge on Kaggle platform, which aims to build predictive models that determine which passengers were more likely to survive using historical data provided by the competition.

Despite the straightforward goal, the dataset presents several obstacles for machine learning approaches, including significant missing values and seemingly irrelevant features. These issues commonly result in reduced predictive accuracy when handled naively. Prevailing beginner strategies focus mainly on key variables such as Age and Sex, which, while important, are insufficient to capture the complexity of survival outcomes.

To address these challenges, this study explores all available features within the Titanic dataset, rather than restricting analysis to a subset. By leveraging comprehensive data processing and feature engineering techniques, this work aims to uncover deeper relationships among variables—potentially revealing influential predictors not initially obvious. Through this approach, the study aspires to improve model performance and provide new insights into factors affecting Titanic survivability.

# 2 Materials and Method

## 2.1 Titanic Dataset

Based on the dataset of passengers who sailed on the Titanic and their survivability, our study evaluated our approach against significant challenges. With 12 variables and 891 observational samples, the dataset was geared towards distinguishing between survived and non-survived passengers on the Titanic. However, there are some major problems. Firstly, nearly 80% of the samples did not have value in Cabin variable, which is important information to detect which part of the ship that passengers have stayed on. Since the bow of the ship sank first, passengers staying at the stern of the ship should have more time to be rescued. Moreover, our examination revealed a lack of meaningful correlation between certain variables within the dataset. This absence of correlation complicated our efforts to identify underlying relationships and patterns, impeding the model’s predictive capabilities.

## 2.2 Problem Overview

.

## 2.3 Proposed Model

.

## 2.4 Implementation Details

.

# 3 Experiment and Results

## 3.1 Experiments Setup

.

## 3.2 Results

.

# 4 Conclusions

.

# References

[1] .

[2] .

[3] .

[4] .

[5] .