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.

October 10, 2025

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.

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. Secondly, there is a class imbalance in this Titanic dataset (approximately 38% survived versus 62% did not, biasing models toward the majority class). 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 Proposed Solution

We tried to fix the following important problems to improve our model results:

- **Handling missing and outlier values**
  Missing or odd values can hurt the model's accuracy by adding noise.

- **Feature engineering**
  Some features show high correlation with each other. Creating new features from such pairs adds more detailed data for the model.

- **Using suitable training algorithm**
  We tested eight algorithms:

  1. Logistic Regression
  2. Decision Tree
  3. Random Forest
  4. Gradient Boosting
  5. AdaBoost
  6. SVC
  7. KNN
  8. Naive Bayes

  The best-performing one was chosen for final model training.

3 Experimental Setup

3.1 Handling Missing Values

Missing data represents a critical issue in the Titanic dataset, with notable proportions observed in both the training and test sets. Specifically, the training set contains missing values for Age (19.87%, 177 instances), Cabin (77.10%, 687 instances), and Embarked (0.22%, 2 instances). Similarly, the test set reports missing values for Age (20.57%, 86 instances), Cabin (78.23%, 327 instances), and Fare (0.24%, 1 instance).

To preserve the dataset's integrity and avoid information loss, an imputation strategy tailored to each feature was adopted:

Missing values in the Age feature were imputed using the mean age of the entire dataset. However, for passengers bearing the title 'Master'—typically those under 12 years old—the mean age of all 'Master' passengers was used. This stratified imputation strategy preserves demographic variability and yields more accurate estimates than applying a global mean to all missing values.

Fare: A single missing value in the test set was imputed with the overall mean fare, ensuring consistency with the fare distribution across the dataset.

Embarked: Given the negligible proportion of missing values (0.22%), the two missing entries in the training set were imputed with the mode value ('S', Southampton), reflecting the most common embarkation port.

Cabin: Due to the extremely high proportion of missing values (77%), the whole feature is removed.

All missing values were detected using isnull().sum() and systematically treated to ensure that no rows were excluded. Post-imputation validation confirmed the absence of null values, guaranteeing a complete dataset for subsequent modeling.

3.2 Feature engineering

To boost predictive power, several features were engineered from the Titanic dataset:

Age bin: Age discretized into six bins $(< 12, 12 - 18, 19 - 25, 26 - 32, 33 - 45, > 60)$ to capture survival trends (e.g., children < 12 had 60% survival rate vs. 38% overall).

Fare bin: Fare values are binned into nine ranges using pd.cut to address skewness and outliers. Higher fares (> 100) correlate with 70% survival, reflecting socio-economic status and lifeboat access.

FamilySize: Calculated as SibSp + Parch + 1 (range: 1–11). Based on the research, FamilySize from 2 to 4 tends to have higher chance of survived, while families that had more than 4 members was harder to survived

FamilyGroup: This feature is based on FamiliSize

Small families (2–4 members) had 55% survival, vs. 30% for alone and 15% for large families (more than 4 members).

3.3 Model

We have chosen eight training algorithms, base on that we validate the test to find the best model for this Titanic dataset. Finally, Gradient Boosting is the best model for this dataset.

Advantages.

- Achieves high accuracy by combining many weak learners.
- Can capture complex non-linear relationships between features.
- Handles both categorical and numerical data effectively.
- Flexible with different loss functions.

Disadvantages.

- Training can be slow due to sequential learning.
- Sensitive to hyperparameter tuning (learning rate, number of trees, depth).
- More prone to overfitting if not regularized properly.
- Less interpretable compared to simple models like Logistic Regression.

4 Experiment and Results

4.1 Experiments

We evaluated several machine learning models on the Titanic dataset. The validation accuracies of the models are reported in Table 1.

Table 1: Validation Accuracy of Different Models

Model	**Validation Accuracy**
Logistic Regression	0.8060
Decision Tree	0.7612
Random Forest	0.8022
Gradient Boosting	**0.8209**
AdaBoost	0.7948
SVC	0.6567
KNN	0.7052
Naive Bayes	0.7799

4.2 Results

Among all models, Gradient Boosting achieved the best performance with an accuracy of 0.8209.

To further evaluate this model, we examined the confusion matrix and the classification report (Table 3).

Table 2: Confusion Matrix (Gradient Boosting)

	Predicted 0	Predicted 1
Actual 0	144	13
Actual 1	35	76

Table 3: Classification Report (Gradient Boosting)

Class	Precision	Recall	F1-score	Support
0 (Not Survived)	0.80	0.92	0.86	157
1 (Survived)	0.85	0.68	0.76	111
Accuracy	0.82			
Macro Avg	0.83	0.80	0.81	268
Weighted Avg	0.82	0.82	0.82	268

5 Conclusions

Despite considerable challenges posed throughout this study, the models developed achieved promising results. After training was completed, out-of-sample evaluation on the test dataset yielded a Kaggle score of 0.78229. While this score does not yet meet the primary research objective, it nevertheless represents substantial progress for the project. Future work will build upon these findings to further enhance model accuracy and robustness.

Table of Contents