Challenge 1:Titanic – Machine Learning from Disaster

Tran Quoc Hoang, Tran Chi Vy,

Phan Tran Huu Tan

Saigon University, Vietnam.

*Corresponding author(s). E-mail(s):

3123580065@sv.sgu.edu.vn;Contributing authors:
3123580015@sv.sgu.edu.vn;

3123580042@sv.sgu.edu.vn;

These authors contributed equally to this work.

Abstract

Titanic is the most well-known maritime catastrophe in the world. As a result, it led to a typical safety management study. Predicting passenger survival is both a challenge and an essential task in risk factor inspection. The value of this study is generated from the implementation of machine learning and data mining methods to predict and understand the patterns between survival and related variables.

The dataset was collected from Kaggle and comprises 12 attributes and 342 survivors. Several valuable columns, such as *Sex* and *Age*, were divided into training and testing subsets. This study explores data mining processes, including exploratory data analysis (EDA), model training, and evaluation using a dataset split. Among all models, the Random Forest achieved the best performance with 82% accuracy, demonstrating that *Gender*, *Age*, and *Class* were the most influential features in the prediction process. This indicates the efficacy of tree-based ensemble methods.

The results provide evidence of the effectiveness of data mining techniques in disaster survival prediction. However, there remain limitations such as small data size and the lack of extensive cross-validation. Future work could improve the output through feature engineering and hyperparameter optimization.

Keywords: data mining, exploratory data analysis, machine learning, maritime catastrophe, classification, survival prediction, Titanic dataset, random forest

1. Introduction

The sinking of RMS Titanic showed a valuable case for safety management. As mentioned in the abstract, predicting individuals who survived a tragic accident is not only historically impressive but also necessary in classification assessing with missing and imbalanced data. The more accurate the model, the more strongly it exactly related in real-life survival.

This study represents the survival prediction task as a supervised binary classification problem using the publicly available Titanic dataset. Key variables include potential attributes (*Sex*, *Age*), socio-economic proxies (*Pclass*, *Fare*), and embarkation attributes (*SibSp*, *Parch*, *Embarked*).

The workflow represents a successful application of exploratory data analysis, imputation, categorical encoding, and a comparative evaluation of classical learning algorithms: Random Forest, Gradient Boosting, Logistic Regression, and k-Nearest Neighbors. Model evaluation enforces not only accuracy but also class-specific diagnostics (precision, recall, F1-score) and confusion matrix execution.

The main contributions of this work are threefold:

1. A reproducible baseline pipeline that highlights outclass predictors of survival in tabular maritime data.
2. An experimental comparison showing that ensemble tree methods outperform linear and distance-based baselines.
3. A small discussion of difficult preprocessing and limitations.

Constraints include limited sample size, lack of extensive cross-validation, and basic feature engineering. Future work should address these through enhanced pipelines, expanded feature construction (e.g., titles, deck extraction, family size), and systematic hyperparameter tuning to improve generalization and minority class detection.

2. Materials and Methods

2.1. Titanic Disaster Dataset

Our discovery forcuses on the Titanic – Machine Learning from Disaster dataset, a interesting binary classification challenge sourced from the Kaggle platform. The dataset is segmented into a training set (train.csv), comprising 891 observational samples, and a test set (test.csv), containing 418 samples. Crucially, the training set incorporates the target variable Survived (0 = Deceased, 1 = Survived), which serves as the basis for model training and testing.

The dataset consists of 12 different variables illustrating passenger characteristics, including socioeconomic class (Pclass), gender (Sex), age (Age), accompanying relatives (SibSp, Parch), and fare (Fare), beside identification variables like Name, Ticket, and Cabin. However, initial data exploration revealed a critical challenge: the presence of enormous missing data across the Age, Cabin, and Embarked columns. This necessitated the implementation of rigorous data cleaning, imputation, and preprocessing procedures to measure the data's quality and appreciation for machine learning model deployment.

2.2. Problem Overview

The objective of this research is to solve a binary classification issue by accurately predicting the survival outcomes of Titanic passengers. The classification is derived from demographic and socio-economic attributes available in the dataset.

Following best practices for structured tabular data, this research prioritized traditional and ensemble machine learning models algorithms such as Gradient Boosting and Random Forest were chosen for their proven ability to capture complex data relationships and achieve high classification accuracy.

However, the performance of these models strongly depends on the quality and interpretability of input features. Consequently, the study enforced enhancing model performance and deadling the dataset's sparse and deverse nature of the data. In addition, this study's result was not the best model in some cases.

2.3 Proposed Model and Feature Engineering

To maximize model performance, the study established a detailed preprocessing pipeline. Data Preprocessing involved Imputation, where Missing Age values were imputed using median by Title, Missing Fare was filled using the dataset median, and Missing Embarked was filled with the mode value. For Encoding and Scaling, Categorical variables (Sex, Embarked, Title, Pclass) were encoded using One-Hot or Label Encoding for model optimization.
Experimental Models: Four machine learning algorithms were evaluated: Logistic Regression, K-Nearest Neighbors (KNN), Random Forest, and Gradient Boosting.

2.4. Implementation Details

2.4.1. Algorithm

Our research explores traditional machine learning models, ranging from simple linear models to more complex ensemble methods, which are applied to our Titanic survival prediction study. Random Forest Model: For tabular data, ensemble methods often provide optimal performance.
We erected a Random Forest classifier with 100 estimators using balanced class weights to handle the imbalanced dataset. This model leverages multiple decision trees and voting mechanisms to make predictions, demonstrating superior classification ability in addressing the survival prediction problem.
Gradient Boosting Model: Gradient Boosting combines weak learners sequentially, with each new model correcting the errors of previous ones. This iterative approach allows the

model to capture complex patterns in the data, making it particularly effective for the Titanic dataset's non-linear relationships between features and survival outcomes.

Logistic Regression Model: As a linear classification model, Logistic Regression provides interpretable results and serves as a baseline for comparison. We implemented it with balanced class weights to address the dataset imbalance, where only 38.38% of passengers survived.

K-Nearest Neighbors (KNN) Model: KNN classifies passengers based on the survival outcomes of their k nearest neighbors in the feature space. We used k=5 neighbors, though this model showed lower performance compared to ensemble methods.

2.4.2. evaluation metrics

In addition to data manipulation through preprocessing techniques, we recognized that applying balance strategies could reduce the effects of data imbalance. Therefore, we experimented with different evaluation metrics to assess model performance. We used confusion matrix for calculating all the metrics, showing true positives, false positives, true negatives, and false negatives for detailed error analysis. Nevertheless, there would be a lack of overall understanding if this process didn't have classification report.

Accuracy Score: The primary metric used to evaluate model performance, calculated as the ratio of correct predictions to total predictions.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision: Measures the proportion of correctly predicted positive instances among all predicted positive instances.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall: Measures the proportion of correctly predicted positive instances among all actual positive instances.

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1-Score: Harmonic mean of precision and recall, providing a balanced measure of model performance.

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 \times TP}{2 \times TP + FP + FN}$$

Note: Where:

TP = True Positives

TN = True Negatives

FP = False Positives

FN = False Negatives

Furthermore, we also intergrate two additional metrics, that are macro average and weighted average. In some situations, especially in case of imbalanced data, the evaluating process would be more sustainable and realistic when macro average and weighted average show their effectiveness.

Macro Average (MAG): Computes the simple arithmetic mean of precision, recall, and F1-score across all classes, treating each class equally regardless of its size.

$$\text{Macro Average} = \frac{1}{N} \sum_{i=1}^{N} M_i$$
Weighted Average (WAG): Calculates the mean of precision, recall, and F1-score across all classes, weighted by the number of samples (support) in each class.
$$\text{Weighted Average} = \frac{\sum_{i=1}^{N} w_i \times M_i}{\sum_{i=1}^{N} w_i}$$
where \(w_i\) represents the number of true instances (support) for class \(i\).

3. Experiment and Results

3.1 Experiments Setup

Dataset and split: Used train.csv (891 samples). The data was split once into train/test sets with an 80/20 split using train_test_split. No cross-validation or hyperparameter was performed
Preprocessing:
Imputation: Filled missing Age and Fare with the median; filled missing Embarked with the mode.
Encoding: Applied label encoding to Sex and Embarked. No feature scaling, one-hot encoding, or engineered features (e.g., Title, FamilySize, IsAlone) were used.
Features used: Pclass, Sex, Age, Fare, SibSp, Parch, Embarked.
Models evaluated: Random Forest, Gradient Boosting, Logistic Regression, and K-Nearest Neighbors. Models were trained on the training split and evaluated on the hold-out test split.
Evaluation metrics: Primary metric was Accuracy on the hold-out set. Additionally, confusion matrices and classification reports (precision/recall/F1) were produced.

3.2. Results

3.2.1 Results on traditional models

We evaluated four algorithms—Random Forest, Gradient Boosting, Logistic Regression, and K-Nearest Neighbors. Accuracy on the hold-out set was used for comparison. Random Forest achieved the highest accuracy (82.12%), followed by Logistic Regression (81.56%), Gradient Boosting (81.01%), and K-Nearest Neighbors (72.07%). The superior performance of the tree-based ensemble relative to the simpler baselines indicates an ability to capture non-linear relationships in the data. Based on these results, Random Forest was selected for the final prediction phase.

3.2.2 Final model performance

The selected Random Forest model achieved 82.12% accuracy on the hold-out validation set, suggesting good generalization under the chosen split. This supports the effectiveness of the applied preprocessing (median/mode imputation and label encoding)

combined with an ensemble learning approach for binary classification on structured tabular data.

3.2.3. Model evaluation visualizations

Biểu đồ so sánh Accuracy giữa các mô hình
Nguồn: Cell 27 (đoạn vẽ results_df barplot)
Caption gợi ý: “Hình 3.2.3-a. So sánh độ chính xác trên tập hold-out giữa bốn mô hình.”
Confusion matrix của mô hình tốt nhất (Random Forest)
Nguồn: Cell 27 (confusion matrix được vẽ trong vòng lặp; chọn hình của “Random Forest”)
Caption gợi ý: “Hình 3.2.3-b. Ma trận nhầm lẫn của Random Forest trên tập hold-out.”
(Khuyến nghị thêm) Feature importance của Random Forest
Nguồn: Thêm một cell mới ngay sau Cell 27 với biểu đồ feature importance (đã gửi code mẫu ở trên)
Caption gợi ý: “Hình 3.2.3-c. Độ quan trọng đặc trưng của Random Forest.”
(Tùy chọn/phụ lục) Phân bố dự đoán submit (0/1)
Nguồn: Cell 30
Caption gợi ý: “Hình Phụ. Phân bố nhãn dự đoán cho bộ test.”

4. Conclusions

The comprehensive evaluation of various machine learning models on the Titanic survival prediction dataset yielded significant and commendable results, confirming the efficacy of preprocessing techniques combined with robust ensemble learning methods.Our research successfully demonstrated that traditional machine learning models, when properly tuned and fed with high-quality, preprocessed features, remain powerful and practical solutions for classification tasks involving structured tabular data. Specifically:Efficacy of Preprocessing: The implementation of rigorous data cleaning, imputation of missing values, and categorical encoding was crucial. These preprocessing steps successfully mitigated the impact of missing data and provided the algorithms with clean, standardized input, significantly boosting the models' predictive power.Superiority of Ensemble Methods: Among the four algorithms evaluated, the Random Forest ensemble method delivered the highest accuracy score of 82.12%. This confirmed that ensemble techniques are exceptionally well-suited for capturing the non-linear decision boundaries inherent in the Titanic dataset, effectively outperforming simpler models like Logistic Regression and K-Nearest Neighbors.Achievement of Performance Goal: By implementing the optimal Random Forest model, the study successfully generated predictions that achieved a competitive accuracy score of 82.12% on the validation data.In conclusion, the study validates a systematic methodology: preprocessing is paramount, and powerful ensemble models are the most effective classifiers for this type of problem. Future work could focus on advanced hyperparameter optimization techniques and feature engineering to further refine the selected models.

5. Acknowledgements

This work was initially developed as part of the Data Mining subject at Saigon University. The authors wish to thank Do Nhu Tai for their guidance.

References

[1] Kaggle, “Titanic – Machine Learning from Disaster,” in *Kaggle Competitions*, 2012. [Online]. Available: https://www.kaggle.com/competitions/titanic

[2] Patil, A., Singh, D.: Predicting survival on the Titanic using machine learning techniques. International Journal of Innovative Research in Computer and Communication Engineering, 5(4), 6429–6435 (2017).

Validation ko có

Ko có onehot ,encoding

Viết công thức latex

Check ngữ pháp blablabla