Survivals of Titanic Prediction Utilizing Tree-based Machine Learning Models

Tianyi Zhao

Department of Communication, University of Colorado Denver, Colorado, United States

Tianyi.2.Zhao@ucdenver.edu

Abstract. The shipwreck of Titanic is a well-known tragedy. Although it happened more than a century ago, researchers are still investigating the patterns of the survivors to gain more insight into human behaviors in catastrophes. This paper adopts machine learning techniques, including decision tree, random forest, and gradient boosting, to conduct a binary classification to predict whether a person survived. The selected models are all tree-based, making it convenient to examine the importance of features. In the preprocessing stage, all numerical features are discretized. This paper first investigates the performances of the models. Subsequently, the model with the best performance generates and studies the importance of the feature. The result demonstrates that the decision tree classifier with a max depth equal to seven achieves the highest accuracy of 0.78. The results of the three models are similar, indicating that the research is robust. The feature importance generated by the decision tree classifier shows that sex and social status significantly impact the survival result. In addition, whether the person is a child also makes a difference. The discretized features do not have enough influence on the result of survival. This paper concludes that the tunned decision tree classifier is the best model to study the features in this paper, but the created features are not effective enough.

Keywords: Supervised learning, feature importance, Titanic.

1. Introduction

The story of Titanic is a famous tragedy. Titanic's collision with an iceberg and its shortage of lifeboats led to the death of 1, 502 out of 2, 224 passengers and crew on April 15, 1912 [1]. Today, researchers are still investigating the patterns in the surviving passengers even a hundred years after the shipwreck. Although it happened more than one hundred years ago, examining who survived the disaster is still helpful to better understand disasters and human behaviors. To understand what characteristics of a person are more likely to increase a person's chances of survival in this tragedy and to provide a reference for future disaster management, the dataset provided by Kaggle is studied using machine learning techniques in this paper.

Multiple researchers have found different algorithms to study the Titanic dataset. Ekinci et al. have compared the performance of twelve different algorithms and concluded that the voting classifier composed of the Gradient Boosting Classifier, K Nearest Neighbors, and Artificial Neural Networks has the highest 0.82 F-measure score among all these twelve models [2]. In another study, Singh A. et al. found that logistic regression has the highest accuracy score compared to naive bayes, decision trees, and random forest classifiers [3]. Another group of researchers, Singh, K. et al. compared the performance of Logistic Regression, K - nearest Neighbours, Support Vector Machines and Decision Tree. Hyperparameter tuning was implemented in this research. They concluded that the decision tree model that they built has the highest accuracy of 93.6% [4].

The researchers have also implemented various kinds of preprocessing techniques. In terms of feature engineering, the techniques seem to be similar. The study by Barhoom et al. constructs a new feature called family size [5], defined as the sum of Parch and Sibsp plus one. Another researcher, Ai, did it slightly differently by turning the feature into a Boolean categorical feature called IsAlone, which is determined by whether there are family members on board. Similarly, continuous numerical features like Fare and Age are binned and encoded [6]. In the same paper of Barhoom [5], the titles are extracted from the name feature to replace the original feature, which is identical to Singh's approach [3].

Researchers have also studied the importance of existing and created features. Cao et al. created various Boolean and categorical features [7], and they found that the Age feature in the dataset has the highest importance, followed by calculated_fare, a feature that they created by dividing fare by family size. They concluded that Age, Title, Sex, and Family Group are the most important factors.

A wide range of machine learning models are implemented in the reviewed works. Research on engineering the numerical features into Boolean features and binned features is conducted. The feature importances of these created features are examined. Yet research that only uses these discretized features has not been found.

This paper aims to fill the gap by analyzing the importance of features after converting all numerical features to categorical ones. For the rest of this paper, the data exploration, data transformation, data engineering, and method introduction will be presented in the methodology section. The results, including the performances of different models and the feature importance, will be presented and discussed in the result & discussion section.

2. Methodology

2.1. Dataset Preparation

The Titanic dataset refers to the dataset posted on Kaggle in this paper [1]. Kaggle provides eleven variables, such as pclass variable for the ticket class, sibsp for the number of siblings and spouses aboard, Parch that stands for the number of parents and children aboard, and Embarked stands for which port they boarded Titanic.

Since the aim is to determine which feature will influence the survival result, the survival result must be predicted. Therefore, the survival column should be treated as the label. This indicates that a binary classification should be conducted as survival is a binary feature. In this paper, plcass, sex, Age, sibsp, Parch, fare and embarked are selected as the features, among which all variables except pclass will be transformed or engineered later. The dataset is already split into a train set and a test set. The former one has 891 entries, and the latter one has 418. The histograms and box chart of the features are shown in Fig. 1 as follows:

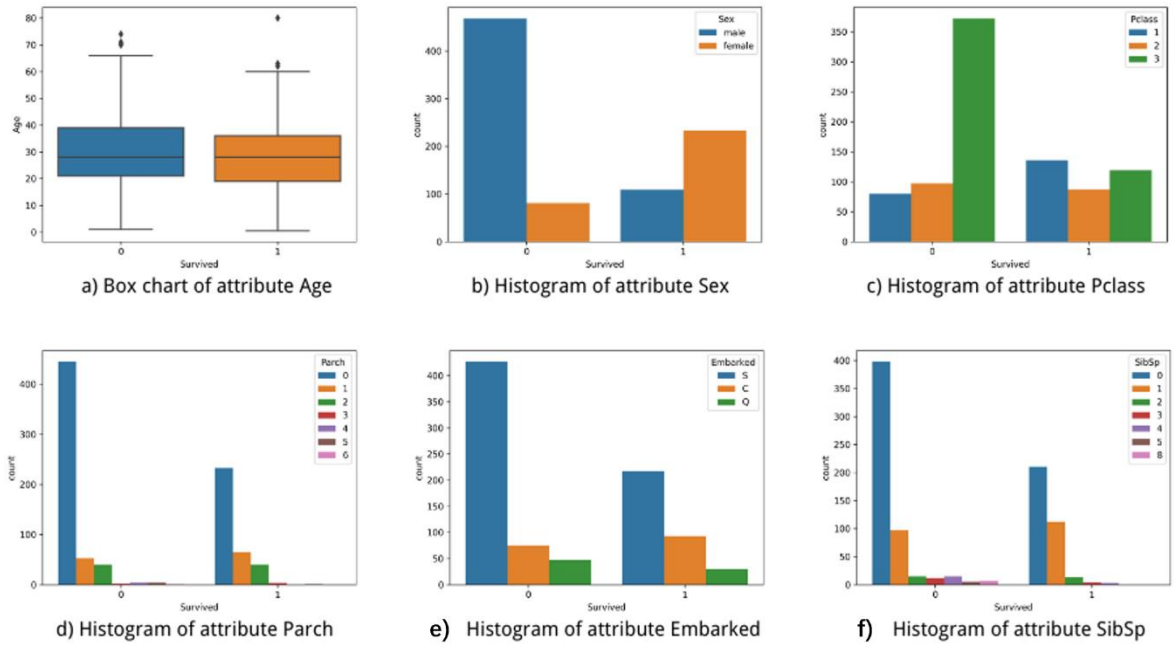

Figure 1. Visualization of features and their relationship with survival (Picture credit: Original)

2.2. Preprocessing

Data preprocessing techniques, including one-hot encoding, discretization, and binarization, were implemented for some features. The Age feature is binarized and altered into the feature IsChild, for which 0 stands for passengers under 12 and 1 for above. Similarly, Sibsp and Parch are summed and binarized into hasFamily. In addition, Fare is binned into fares_cat, which contains fare levels from one to four, corresponding to low to high fare values. Finally, Embarked and sex are one hot encoded.

Therefore, the input features are Pclass, IsChild, hasFamily, fares_cat, female, male, Embarked_C, Embarked_Q, and Embarked_S. The normalization is not implemented since the features are all binned, binarized, and one hot encoded. Also, the models used in this paper are invariant to the features' scale.

2.3. Machine Learning Models

2.3.1 Decision Tree

Decision tree is a widely used method for classification tasks, it has a tree-like branch structure. It includes the root node, some internal nodes, and leaf nodes. The root and each internal node represent a comparison condition based on values of some input features. Each of these nodes is connected to two or more child nodes, which are treated as the result of the current decision and the condition of the decision in the next step. Decisions are made at every node, from the root node to the internal nodes at every level. Finally, they reach the leaf nodes, which represent the result of classification. During this process, a sequence of decisions based on the input features is made to determine the classification result.

2.3.2 Random Forest

Random forest is an ensemble classification method. It is composed of many individual decision trees. This method randomly selects features as inputs to build individual trees and each split within a tree. It also does bootstrap, which means that only some random samples from the train are set by sampling with replacement so that the visible data to each tree is different. The output classification result of each of these decision trees serves as the input of a voting system. The voting system then outputs the mode of these classification results, which is also the final output of the random forest classifier.

2.3.3 Gradient Boosting

Gradient Boosting builds multiple tree-based structures sequentially. The initial prediction is typically the logarithm of the odds of the positive class occurrences over the negative ones, which is then converted into a probability to establish a starting point. Then, a large number of trees with limited leaves are built sequentially based on the residuals of all previous trees to correct the previous errors. The results of each individual tree are scaled by a learning rate before it is aggregated and then added. Finally, A logistic function is applied to convert the combined predictions into a probability.

3. Result and Discussion

3.1. Model Performance

Since the testing set of Kaggle competition datasets is unavailable, a validation set is created to evaluate the model performances. Ten-fold cross-validations are conducted on these three models. The models' performances are evaluated on each validation fold. Finally, the average accuracy score across all ten folds on each individual model is used to reflect each model's performance on the validation set.

The accuracies of the three models on the validation set after hyperparameter tunning using random search are presented in the following Table 1:

Table 1. Performance of the models

Model Name	Accuracy (validation set)	Accuracy (testing set)
Gradient Boosting Classifier	0.8250	0.7703
Decision Tree Classifier	0.8193	0.7728
Random Forest Classifier	0.8216	0.7703

The accuracy score shows that the gradient boosting classifier with a learning rate of 0.1785, a max depth of 5, and 144 estimators has achieved the highest accuracy score on the validation set among the three models. However, the decision tree with a max depth equal to 7 yields a higher accuracy score on the testing set.

The results of the three models are almost identical. One potential cause of this result is that these three models are all inherently similar since gradient boosting and random forest are ensembles of decision trees. The similar accuracy score also indicates that the finding is likely robust. The accuracy of 77% is lower than the models in previous studies that discretized some of the features. For example, there is a random forest classifier with an accuracy of 83% in the study of Cao et al. and an SVM classifier with an accuracy of 82.82% in Ai's work [6]. This suggests that applying discretization on every feature may not be helpful on the Titanic dataset. Future work should be done to compare the performances of the models trained on continuous and discretized features that adopt the same data preprocessing process.

3.2. Feature Importance

The feature importances are provided by the decision tree model, which has the best performance on the testing set. The feature importance is visualized in the following Fig. 2:

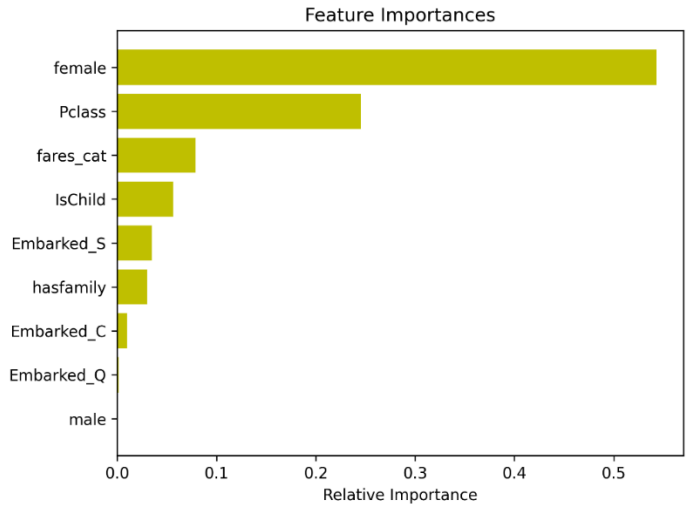

Figure 2. The feature importance (Picture credit: Original)

The importance of the features revealed that the discretized features are less important to the prediction result. This result coincides with general results from previous studies. The high relative importance of female, Pclass, and fares_cat features suggest the impact of gender and ticket price on the result of survival. The latter is a reflection of the social status of passengers. This might mean that being a woman and being in a higher social class will likely survive. The importance of IsChild indicates that children have greater chances of survival in such disasters. However, this feature does not drastically impact survival.

In terms of limitations, this paper does not sufficiently explore the discretization of fares and Age. Hence, in future work, more methods of discretizing fares and Age features should be investigated to gain more insights into survival patterns. In addition, more advanced models e.g. neural networks can be considered due to their excellent performance in many tasks [8-10].

4. Conclusion

The goal of machine learning in this study is to examine what features influence the result of survival on Titanic. This paper explores the model performance and the importance of features in the Titanic dataset after discretizing all numerical features. The models implemented in this study are the Decision Tree classifier, Random Forest classifier, and Gradient Boosting classifier. The tunned Gradient Boosting classifier performs slightly better than the other two models. In addition, the discretized features only make limited impacts on the prediction result. This paper does not compare the importance of the original features and the features created by discretizing them. Moreover, the methods of discretizing fares the Age feature have not been sufficiently explored. Future work should be done to overcome these limitations.

References

[1] Kaggle. Titanic - Machine Learning from Disaster. Kaggle. https://www.kaggle.com/c/titanic/overview, 2024.

[2] Ekinci E, Omurca S İ, Acun N A. Comparative study on machine learning techniques using Titanic dataset. 7th international conference on advanced technologies, 2018, 411-416.

[3] Singh A, Saraswat S, Faujdar N. Analyzing Titanic disaster using machine learning algorithms. 2017 International Conference on Computing, Communication and Automation (ICCCA 2017), 2017, 406-411. Doi: 10.1109/CCAA.2017.8229835.

[4] Singh K, Nagpal R, Sehgal R. Exploratory Data Analysis and Machine Learning on Titanic Disaster Dataset, 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2020, 320-326. Doi: 10.1109/Confluence47617.2020.9057955.

[5] Barhoom A M, Khalil A J, Abu-Nasser B S, Musleh M M, Naser S S A. Predicting Titanic Survivors using Artificial Neural Network. International Journal of Academic Engineering Research, 2019, 3 (9): 8-12.

[6] Ai Y. Predicting Titanic survivors by using machine learning. Highlights in Science Engineering and Technology, 2023, 34, 360–367.

[7] Cao Y, Xie W, Dong C, Qiu J. Titanic Machine Learning Study from Disaster. Applied Economics & Statistics Research Report, University of Delaware, 2020, RR20-01.

[8] Qiu Y, Wang J, Jin Z, Chen H, Zhang M, Guo L. Pose-guided matching based on deep learning for assessing quality of action on rehabilitation training. Biomedical Signal Processing and Control. 2022 1; 72: 103323.

[9] Woźniak M, Wieczorek M, Siłka J. BiLSTM deep neural network model for imbalanced medical data of IoT systems. Future Generation Computer Systems. 2023 Apr 1; 141: 489-99.

[10] Ding Y, Zhang Z, Zhao X, Hong D, Cai W, Yang N, Wang B. Multi-scale receptive fields: Graph attention neural network for hyperspectral image classification. Expert Systems with Applications. 2023 Aug 1; 223: 119858.