Comparative Analysis of Models Based on Titanic Survival Predictions

Yulin Huang

Liwuili High School, Ningbo, Zhejiang, 315041, China
tangyicheng@xdf.cn

Abstract. Survival in natural disasters and major accidents is difficult to predict, and in many cases, it is difficult to extrapolate. In this paper, and the probability of survival of the Titanic tourists is discussed. First of all, the characteristics of the passengers in the Titanic data collection from the Kaggle, data processing and data analysis. In data processing, the method of taking the median is used to fill the vacancy of data, making the result become more accurate. Through Decision trees, Random Forest and the general survival prediction model experiments, the most suitable model was analyzed and its highest accuracy and accuracy was determined. It is found that the random forest model has high precision and accuracy in the application of survival prediction. The correlation survival prediction model has been applied to the survival prediction of natural disasters such as ship accidents at sea. Experiments show that the correlation model is reliable and effective, and the prediction model of survival rate is effective.

Keywords: Titanic; decision tree; random forest; survival prediction model; accident rescue

1 Introduction

1.1 Background

During the night of April 14, 1912, cruising on the dark ocean surface, the luxurious RMS titanic, a cruise ship, collided with an iceberg. Two hours and 40 minutes later it sank on its maiden voyage, resulting in the death of 1,501 people--more than two-thirds of its 2,207 passengers and crew. The sinking astonished the world and is still considered one of the deadliest maritime disasters in history, as well as one of the most infamous. The ship's 'unsinkable' status and the most advanced maritime technology on board made it a severe disaster at the time of the disaster. On that terrible night, many classes of individuals of various ages and genders were present; the unfortunate thing was that there weren't many lifeboats available for rescue. Many men who had been replaced by the numerous women and children on board were among the deceased. The folks who were traveling in the second class had already passed away [1].

Y. Wang (ed.), *Proceedings of the 2024 2nd International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2024)*, Advances in Computer Science Research 115,
https://doi.org/10.2991/978-94-6463-540-9_17

1.2 Related research

It was soon discovered that during the sinking of the ship, a very remarkable phenomenon was discovered. More than 70% of the women and children were rescued because of the order to put women and children first. Even in life-threatening situations, people tend to act in ways that are socially expected, rather than rationally or selfishly, which illustrates the power of internalized social scale. The primary emphasis on the Titanic was placed on social norms and social status [2]. In shipwrecks, there is an element of luck among the survivors, but it is not entirely random who survives and who dies.

The researchers applied machine learning algorithms to predict whether the sinking of the Titanic was a tourist, and the ticket prices, age, gender, nationality, cabin space and other functions will be used to predict. The process of predictive analysis involves the utilization of computational methods to discern significant and valuable patterns within large datasets, employing machine learning algorithms to forecast the survival rates of various feature combinations [3].

With the continuous development of the industrial level in modern society, more and more scholars have adopted different algorithm models for different applicable situations. Including multivariate regression model, neural network model, decision tree model [4], gray theory model [5], integrated learning algorithm model [6], logistic regression model, random forest model [7], Ada boost-CART model [8], etc.

1.3 Objection

The purpose of this paper is to study and mine the variety of information in the available datasets and through various data in the field of survival. Domain application analysis to understand the impact of each domain on passenger survival. A machine learning technique is used to make the predictions for more recent data sets. The accuracy of the data analysis using the implemented algorithms will be examined. The most accurate algorithm is recommended for predictions after it has been evaluated with other algorithms based on accuracy.

2 Methodology

2.1 Data

The Titanic data set contains personal information and the survival of some of the passengers and crew members of the Titanic when it hit an iceberg in 1912. This data shows the identity of the passenger including name, age, gender, how many brothers and sisters, father and son, ticket price, ship number, cabin number, departure port and so on (Table 1). Because the model needs to be built to predict the test set, more people choose to find the missing value and fill in the missing value before building the model, and then analyze the data to improve the data structure.

Table 1. Input data set

Name	chr	“Braund, Mr.Owen Harris” …
Sex	chr	“female” “male” “male” …
Age	num	26 33 39 51 3 25 19 …
Passenger ID	num	1 2 3 4 5 6 7 8 9 10 …
Pclass	num	3 3 3 1 3 1 3 1 3 2 …
Embarked	chr	“S” “C” “C” “S” …
Ticket	chr	“A/5 21381” “PC 17689” …
Parch	num	0 0 1 0 0 0 0 1 2 0 …
Cabin	chr	“C63” “C147” “C49” …
Fare	num	7.26 73.28 7.52 57.1 8.35 …
SipSp	num	1 1 0 1 0 0 0 3 0 1 …
Survived	num	0 1 0 1 1 0 0 0 1 …

2.2 Data processing

As shown in Figure 1, the analysis of the data can be divided into three steps. The first step is the data filling. Some data may have gaps, so the filling of data plays a very important role. For example, using the median to fill the data can make the data more complete. The second step is data cleaning, including the judgment of the nature of the data and analysis of the data. The third step is data modeling, through data modeling can better analyze the structure of data, and better understand the data. It is possible through these three steps. Better overall analysis of the data.

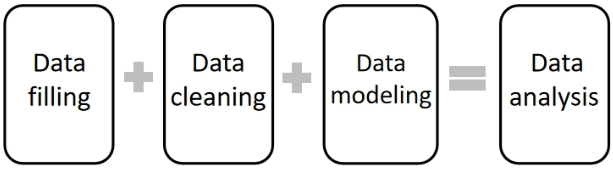


Figure 1. Data analysis

2.3 Model chooses

There is a methodical process for selecting a specific model for the given situation following data analysis [9]. A particular machine learning model is required to solve the challenge. Figure 2 shows the main flow of fitting a mechanical learning model. First simplifying the problem, and then finding the appropriate model to fit it. In the process of fitting the model, find a model that has more accurate prediction results to predict, and finally decide which way to predict.

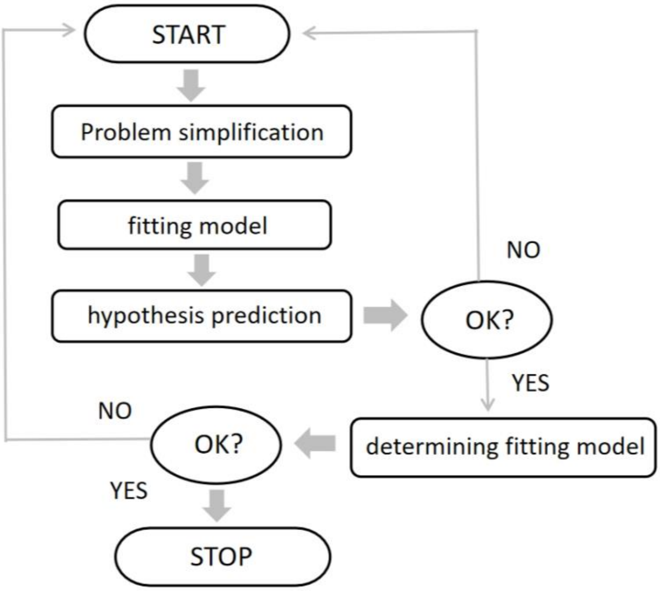


Figure 2. Process of Fitting a Machine Learning Model

The models to be used in the next study, such as the decision tree model, random forest model, and multiple regression model, are determined through these steps. Choose the best model from these models to forecast the end result. The range of considerations chosen includes the accuracy and practicality of the forecast.

3 Results

3.1 Data analysis

After the data is processed and the model is fitted, the data will be input and classified. There are many options for importing datasets. For example, data can be entered into R studio for processing.

The dependent variable in these Titanic data is whether someone has survived. There were 1,300 passengers and 886 civil servants among the 2,186 individuals on board. Passengers are the main subject of practical research. Of the passengers, 840 were men, or 65 percent of the total, and 43 were civil servants. 460 of the 1,300 passengers, or 35% of the total, were female. With just 20 lifeboats on board, the Titanic could accommodate up to 1,178 passengers, or 52% of the total number on board, making lifeboats a rare commodity. It is not difficult to see that the survival rate of more than 1,500 tourists reached 38%. Through the prediction of the position class, we can find that the survival rate of the first class is 42%, and the survival rate of the second class is 24%. From the gender can be found that the survival rate of women reached 67.8%, higher than that of men 32.2%, it can be said that the survival probability of women is far greater than that of men.

In terms of age, adolescents account for a large proportion of the survivors, accounting for 66.3 percent and 14 percent respectively, more than three-quarters, while the probability of survival rate of the elderly is the lowest.

It's not difficult to find it through the analysis of ticket prices again. Tickets that cost more than 100 are available in the first, second, and third classes. The cost of the ticket is less than 50. The most common number of survivors is under 50, but the base is large and the probability of survival is low. The 500-550 range has a survival rate that can be as high as 100%.

Due to more time on the Titanic, the lifeboat restrictions were relaxed, leading to a higher probability of survival in first-class and second-class cabins. First- and second-class passengers may benefit from this, but not third-class passengers. First-class passengers are able to gain the advantage of information, including access to critical lifesaving information, because crews are more inclined to take care of the rich and powerful. First-class and second-class passengers have a higher probability of survival than third-class and lower-class passengers due to their faster escape.

The number of lifeboats on the Titanic was limited in the proportion of men and women, and in the face of such a large number they chose to rescue women first. The first is the relatively young teenagers, and then taking into account women, children and other people who need to be taken care of. So men leave more chances for women to survive, so the probability of survival rate of women is much higher than that of men, and the probability of survival of teenagers is much higher than that of middle-aged people and elderly people.

The probability of survival rate on the Titanic is analyzed from several different perspectives, such as age, gender, personal details of passengers, and the relationship between the number of seats and the number of tickets.

3.2 Model training

A probabilistic model of the survival probability for a typical passenger was employed as an analytic method,

$$\Pr\left(y = 1 \mid x_1, x_2, \dots, x_k\right) = \Phi(\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k) \tag{1}$$

A dummy variable called Y in this equation indicates whether the passenger is alive (Y=1) or not (Y=0). The factors $(x_1, x_2, \dots, x_k)$, including sex, age, ticket price, and so on, are all explanatory variables in this formula. The generation's estimates are contained in the parameters $(\alpha, \beta_1 x_1, \beta_2 x_2, \cdots, \beta_k x_k)$. is the normal distribution function with cumulative standard form. Maintaining the probability Pr (y=1) is the function; the value range is 0-1, and each passenger adds an observation value to it. $(x_1, x_2, \dots, x_k)$ The maximum likelihood approach can be used to estimate the parameters from a sample of these observations that are thought to be independent. A typical probabilistic model is this one [10,11].

$$\frac{\partial \Pr\left(y = 1 \mid x_1, x_2, \ldots, x_k\right)}{\partial x_j} = \beta_j \phi(\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k) \tag{2}$$

Since Φ> 0, the standard normal density function (Φ) is used instead of the cumulative density (Φ). The marginal effect's sign is the same as that of, and in the case of a discrete, a difference is employed in place of a partial derivative. Tourist survival probability is modeled in a very traditional and common manner.

In addition, this paper creates a decision tree model to construct a CART decision tree, and pruning the decision tree. Considering all the data points and generating a complex tree, there may be a fitting situation, the more complex the number of decisions, the higher the degree of overfitting. It can also create a decision tree model to better and more intuitively predict the survival rate of its visitors.

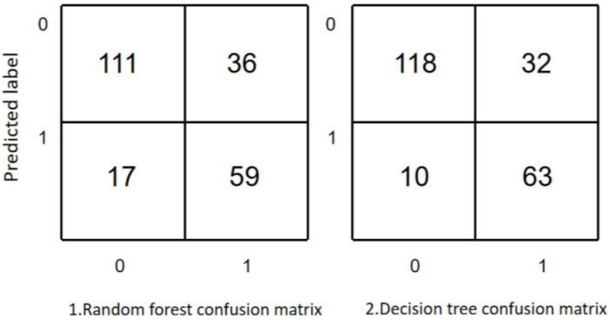
The decision tree model only gets a decision tree to predict. If multiple decision trees are used to train, and then use the crowd of the decision trees to get the prediction results, this is the algorithm of random forest. Since random forest model predictions do not use a single decision tree but use the vote of all decision trees for classification, the overfitting problem of a single decision tree does not affect the final prediction results.

3.3 Model evaluation

The first is to try the general model, the results show that its prediction accuracy is between 0.76 and 0.79, and we can find that the general prediction model has limitations and singularity. Therefore, the prediction and calculation of other models are carried out.

It is not difficult to find from the calculation and prediction, compared with the decision tree model confusion matrix, that random forests in two categories have better prediction effects and higher accuracy. Figure 3 illustrates this point by comparing the confusion matrix data of the random forest and decision tree models. The random forest model's prediction data set is shown to be more efficient, indicating stronger applicability and accuracy. From the data, it can be found that the accuracy of the prediction in the decision number model is between 0.78 and 0.82. In the prediction of the random forest model, the accuracy is about 0.8 to 0.86, The data suggests that the random forest model performs better in predictions than the decision tree model and that its predictions are more accurate.

The superiority of the random forest model over the decision tree model is shown. As a result, when predicting deeper decision trees, the random forest model can predict data sets with greater efficiency. The choice of a random forest model is typically made by researchers to predict the survival of various disasters and difficulties.



Predicted label	0	1
0	111	36
1	17	59

1.Random forest confusion matrix

	0	1
0	118	32
1	10	63

2.Decision tree confusion matrix

Figure 3. Visual confusion matrix

3.4 Discussion

It can also use the Ada boost model. The model algorithm in common with the random forest is that a single learner can be a model of a decision tree. However, the use method of combining the model is not the same, namely, the Ada boost model is represented by using a linear combination of the input from the individual learner. The prediction accuracy of this model for the survival probability problem is as high as 0.86.

From the above several methods, it can be concluded that different methods have different results for probability prediction, such as the accuracy of random forest is higher than the accuracy of decision number. So in the probability of predicting different disasters or different probability of survival rates, you will choose a more accurate model to predict.

Through the application of the above models, the survival rate of the Titanic can be predicted.

4 Conclusion

In this paper, the survival prediction model is constructed to predict the survival of Titanic tourists. First of all, data collection, data processing and data analysis, and the data set into the relevant program, and then use the decision tree model, random forest model, Ada boost model and the general survival prediction model to analyze the Titanic data, and predict the survival. The random forest model is investigated and it is discovered that there are no overfitting or underfitting issues and that the model exhibits a high accuracy. Finally, the survival prediction of the Ada boost model is proved to be better than that of the decision tree. The survival prediction model can predict the survival of people in various natural disasters and different difficulties, and carry out corresponding rescue measures according to the prediction results, so as to improve the efficiency of natural disaster accident rescue.

However, this paper only aimed at four models to predict, data feature selection is limited, and the accuracy of the model still needs to be improved. In the future, more predictive models will be designed, and data features will be added, data sets and different data will be added, and multiple factors will be considered to make more accurate and scientific predictions.

References

1. Analyzing Titanic disaster using machine learning algorithms-Computing, Communication and Automation (ICCCA), 2017 International Conference on 21 December 2017, IEEE.
2. [2] Interaction of natural survival instincts and internalized social norms exploring the Titanic and Lusitania disasters, William J. Baumol, New York University, New York, NY, and approved January 21, 2010
3. [3] Design of the Titanic survival prediction model, Ao Hua Jian, Yu Kaichao, School of Mechanical and Electrical Engineering, Kunming University of Science and Technology, Kunming 650093
4. [4] Yang Z H, Bai J W and Chen Z H. Object-oriented feature wetland decision tree classification method using Sentinel-2A images [J]. Journal of Surveying and Mapping Science and Technology, 2019, 36 (3): 262-268.
5. [5] Li L H, Zhu J S and Xu Y. Study on the spatial distribution prediction method of passenger flow in Beijing-Shanghai High-speed Railway [J]. Railway Transportation and Economy, 2017, 39(6): 32-36.
6. [6] Zhang F, Wu Y Q. Forecdiction of railway logistics demand based on GB-CART integration algorithm [J]. World Science and Technology Research and Development, 2016, 38 (6): 1311-1314.
7. [7] Han J J, Nan S W and Li J P. Research on the Prediction and Control of Grain reactor Mechanical ventilation Temperature based on Random Forest Algorithm [J]. Journal of Henan University of Technology: Natural Science Edition, 2019, 40 (5): 107-113.
8. [8] Wang Y, Fang W and Wang L. Train load factor prediction based on Adaboost-CART model [J]. China Railway, 2019, (10): 34-38
9. [9] Yogesh Kakde, Shefali Agrawal, Predicting Survival on Titanic by Applying Exploratory Data Analytics and Machine Learning Techniques, International Journal of Computer Applications (0975 – 8887)
10. [10] Baum CF (2003) An Introduction to Modern Econometrics Using Stata (Stata Press, College Station, TX).
11. [11] Woodldridge JM (2002) Econometric Analysis of Cross-Section and Panel Data (MIT Press, Cambridge, MA).