Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 5
AIDML 2024

Survival prediction and analysis of Titanic based on logistic regression and KNN

Jiale Li*

Henan experimental high school, Henan, China

* Corresponding Author: 1707060215@stu.hrbust.edu.cn

Abstract. The main purpose of this paper is to use machine learning methods to identify factors related to the survival of Titanic passengers and analyze model parameters. The study first speculated on the survival factors of Titanic passengers and conducted data visualization tests on these speculations. Subsequently, this article cleaned up the relevant data, abandoned the factor of excessive missing data, made up for a small amount of missing data, and then constructed a relevant machine model. Finally, by comparing the accuracy of the two models, a more efficient and accurate model was selected. Through relevant work, it has been concluded that the logistic regression model is more effective than the K-nearest neighbor (KNN) model and that passenger age, economic status, and cabin class are closely related to passenger survival. This article provides valuable references for relevant experimental research by introducing machine learning methods for effective survival prediction. In the future, research will further delve into the methods and solutions of deep learning.

Keywords: Survival of Titanic passengers; Logistic regression model; Machine learning.

1. Introduction

On April 15, 1912, the Titanic cruise ship struck an iceberg, resulting in the deaths of approximately 1,500 passengers and crew. The deadly incident is still forcing researchers and analysts to figure out exactly what factors caused some passengers to survive while others died. The Titanic disaster is one of the most famous shipwrecks in world history. This is a British cruise ship that sank in the North Atlantic just hours after colliding with a giant iceberg. While there are facts to support the cause of the sinking, there is a variety of speculation about the survival rate of passengers in the Titanic disaster. [1]

Eric Lam and Tang used the Titanic problem to compare three different algorithms-naive Bayes, decision tree analysis, and support vector machine-and concluded that gender is the main feature that accurately predicts survival. Among other things, they believe that selecting important features is crucial to achieving better results. There was no significant difference in accuracy between the three methods they used [2]. The predictions are processed using the random forest and decision tree algorithms used by Trevor Stephens, who uses several parameters. However, he did not mention the accuracy of the implemented algorithm. [3] Bruno S. Frey, David A. Savage, and Benno Tzogler concluded that young people had lower mortality rates than older people, and that given their economic status, first-class passengers on the ship were more likely to save themselves than third-class passengers on the Titanic. [4]

The primary goal of the study is to analyze the Titanic accident by using various machine learning algorithms to determine the correlation between passenger survival and passenger characteristics. In particular, this work compares algorithms based on percentages of accuracy on test data sets. In the study, survival estimates for the Titanic are related to each passenger's name, sex, age, class of cabin, number of siblings or spouses on board, number of parents or children, class of cabin, ferry ticket number, fare, and boarding conditions. Specifically, first, exploratory data analysis. Explore the variety of information in the available data set and make sense of it by applying the relevant data, second, using machine learning algorithms and checking the algorithms and accuracy, and third, coming up with the best predictive model by comparing the accuracy of the two algorithms. The

experimental results show that the accuracy of the logistic regression model is higher than that of the KNN model.

2. Methodology

2.1. Dataset Description and Preprocessing

The dataset used in this study called Titanic Survival Predictions is sourced from Kaggle. [5] The data consists of 419 rows in the train, which is a sample of passengers with relevant labels, classifying the survival of tourists in different situations, which can more intuitively show the survival results of tourists on board. Each passenger's name, sex, age, cabin class, as well as the number of siblings or spouses on board, the number of parents or children, cabin, ticket number, fare, and boarding status. After that, in addition to having data sets, it is necessary to check whether these so-called data sets are complete and predict which factors are relevant to Titanic survival predictions. Therefore, many factors can be associated with the prediction such as gender, age, cabin class, number of siblings or spouses on board, number of parents or children, cabin class, ferry ticket number, fare, etc. In this process, it is observed that there are a total of 891 passengers on this ship, and there are three factors with missing values, including cabin, embarked, and age, while the missing values of the other two factors can be filled except the cabin is unfillable. In short, data played a huge role in this process, and the data was presented in different ways, such as charts, etc. to make it more intuitive to observe each situation relevant to the experiment, such as the factors related to the survival of passengers on the Titanic.

2.2. Proposed Approach

The primary goal of the study is to analyze the Titanic disaster using various machine learning algorithms to ensure that the correlation between passenger survival and passenger characteristics is determined. In particular, this work compares algorithms based on the percentage accuracy of test data sets. In the study, survival rates on the Titanic are linked to a variety of factors, such as each passenger's name, sex, age, class of cabin, number of siblings or spouses on board, number of parents or children, class of cabin, ticket number, fare, and boarding conditions. Specifically, first, conducting exploratory data analysis, exploring the variety of information in the available data set and making sense of it by applying relevant data; Secondly, using machine learning algorithms and checking the algorithms and accuracy; Third, the best prediction model is proposed by comparing the accuracy of the two algorithms. The process is shown in the Figure 1.

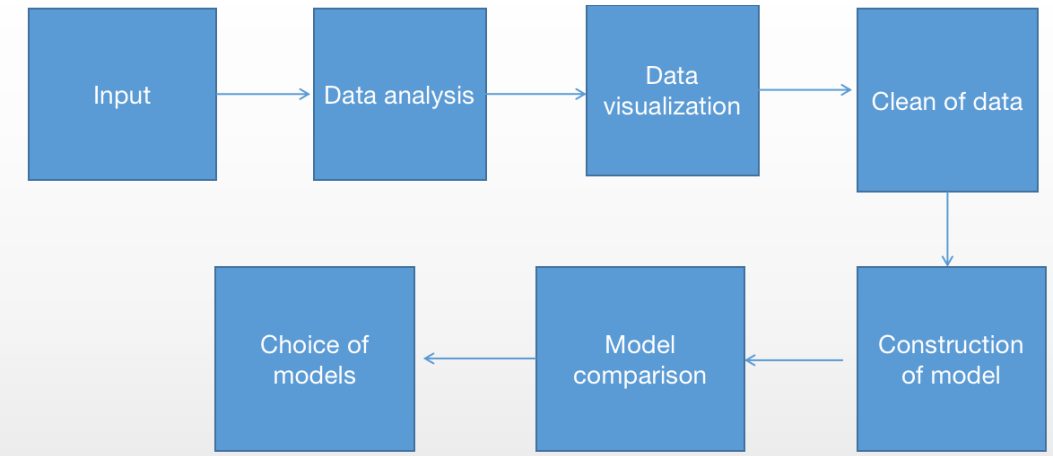

Figure 1. The pipeline of the model

2.2.1. Logistic regression

The logistic regression is commonly used for classification and predictive analysis [6]. Logistic regression estimates the probability of an event occurring, such as voting or not voting, based on a given data set of independent variables. Since the outcome is a probability, the range of the dependent

variable is 0 and 1. In logistic regression, the Logit transformation is applied to the odds, which is the probability of success divided by the probability of failure. This is also commonly referred to as logarithmic odds or the natural logarithm of odds, and this logical function is represented by the following:

$$Logit(pi) = \frac{1}{(1+exp(-pi))} \quad (1)$$

where logit(pi) is the dependent or response variable and x is the independent variable. Maximum likelihood estimation (MLE) serves as a pivotal tool in estimating the parameters (β) of a model. This technique iteratively explores various beta values to achieve the optimal fit of the logarithmic ratio. Each iteration contributes to the generation of a logarithmic likelihood function, which is then maximized by logistic regression to pinpoint the best parameter estimate. Once the optimal coefficient (or coefficients for models with multiple independent variables) is identified, the conditional probability of each observation is calculated, recorded, and aggregated to produce a prediction probability. For binary classification, probabilities less than 0.5 are predicted as 0, while those greater than or equal to 0.5 are predicted as 1. Following model calculation, it is advisable to assess the model's predictive performance on the dependent variable, a practice known as goodness of fit evaluation.

The steps to evaluate accuracy using Logistic regression are as follows: Step 1: Read the data set. Step 2: Ensure its factors. Step 3: Check for missing values. Therefore, once the missing value is determined, attributes that are not relevant to the decision can be removed. In the Titanic dataset, tickets, cabins, names, and passenger ID cards were not used to analyze viability. Therefore, delete these properties. Step 4: Approximate the age of the traveler according to the respective rank. This essentially means that people with the first type of respect are more likely to survive than those with the second. Step 5: Change the unmitigated factor to a false pointer. Step 6: Evaluate the model. Step 7: Finally calculate the accuracy. For our dataset, the accuracy rate was 79.19 percent.

2.2.2. K-nearest neighbor (KNN)

The KNN model is one of the most common and simplest nonparametric classification algorithms, and it belongs to the same classification algorithm as logistic regression. However, unlike logistic regression, the KNN does not make any assumptions about the model and does not require training, which is what logistic regression models need to do. KNN can be used not only for classification, but also for regression. It uses a distance metric to measure the degree of similarity between the training sample and the test sample and assigns the test sample to the k classes closest to the training sample. In terms of proximity, the KNN model is mainly based on Euclidean distance [8].

However, even though the KNN model has become one of the most popular and widely used methods due to its high accuracy, it inevitably still has many shortcomings [9]. For example, because for each text to be classified, the KNN model must calculate its distance to all known samples to obtain its k nearest neighbor points, this also causes KNN to be computationally heavy. A common way to solve this problem is to edit the known sample points in advance and delete the samples that have little impact on the classification. This algorithm is more suitable for automatic classification of class domains with large sample sizes, while it is easy to misclassify class domains with small sample sizes.

3. Results and Discussion

This chapter is a summary of the experiment, which describes the visual data and the performance of different models. Therefore, this chapter analyzes each of the factors that affect the survival of Titanic passengers in the numerous visual data, such as age, gender, cabin, etc. In addition to these, there are also performance comparisons between different models.

Table1. Survival rate of different factors

Sex	Pclass	Cabin
Percentage of females who survived: 74.2 Percentage of males who survived: 18.9	Percentage of Pclass = 1 who survived: 62.96 Percentage of Pclass = 2 who survived: 47.28 Percentage of Pclass = 3 who survived: 24.24	Percentage of CabinBool = 1 who survived: 66.7 Percentage of CabinBool = 0 who survived: 30.0

Table 1 depicts three of the factors that were associated with the survival rate of Titanic passengers. It can be seen from the table that in terms of gender, the survival rate of females is higher than that of males, which is 74.2 and 18.9 respectively. This is because when Titanic sank, according to the principle of "women and children first", the lifeboats gave priority to females and children on board, which resulted in a particularly prominent female birth rate. Second, the second column of Table 1 shows that people with higher social status in society are more likely to survive, so their survival rates are 62.9, 47.3, and 27.2 respectively. This is because in an emergency, the first class is usually allowed to board the lifeboat first, which gives them a greater chance to be rescued. Third, the third section of Table1 shows that the people whose cabin numbers are recorded are more likely to survive, so their survival rates are 66.7 and 30.0 respectively, which is because the passengers whose cabin numbers are recorded may be the ship's staff or crew, who usually have more knowledge and resources to protect themselves and are more likely to survive in a disaster. Cabin numbers may also be linked to promoted social status and wealth, and some promotions with high social status and wealth may have easier access to lifeboats and other resources, increasing their chances of survival.

Some of the results of this data visualization show some of the factors that affect passenger survival rates and their probabilities to help users understand the data more carefully.

Table2. The accuracy of different models

Model	Score
Logical regression	79.19
KNN	77.66

From Table 2, it can be seen that the accuracy of the logistic regression model is 1.53 higher than that of KNN[10]. There are many reasons for this, which can be divided into four main reasons: Data distribution: Logistic regression models assume that the data follows a specific probability distribution, usually a Gaussian distribution, while KNN models do not assume a data distribution. If the true distribution of the data matches the assumptions of the logistic regression model, then this model may be more accurate. Feature selection: The logistic regression model optimizes the weight of features during the modeling process, while the KNN model does not significantly optimize the importance of features. If the selected feature set logic is more differentiated from the regression model, the logistic regression model may perform better. Parameter adjustment: The K value in the KNN model needs to be adjusted, and different K values may result in different accuracy. If an inappropriate K value is selected, the performance of the KNN model may be affected. Data volume: The KNN model may not perform well when the data volume increases, because the distance between samples needs to be calculated and the computational complexity is higher. Logistic regression models generally have better performance on large data sets.

All in all, the difference in accuracy selection may be caused by many factors such as data characteristics, model, parameter adjustment, etc., which need to be analyzed on a case-by-case basis.

4. Conclusion

The main purpose of this paper is to use machine learning methods to identify factors related to Titanic passenger survival and test which machine learning model is more accurate. This experiment first speculated on the survival factors of Titanic passengers, and conducted data visualization tests on these guesses, then cleaned the relevant data, abandoned the factors with too much missing data, made up the small amount of missing data, and then built the relevant machine model. Finally, by comparing the accuracy of the two models, a more efficient and accurate model was selected. Through relevant work, it is concluded that the logistic regression model can play a better role than the KNN model, and it is concluded that passenger age, economic status, and cabin class of passengers are closely related to the survival of passengers. In the future, it can be considered to introduce some other kinds of machine models such as Support vector machines to solve relevant problems rather than just logistic regression and other classification models.

References

[1] S. Aakriti, S. Saraswat, and N. Faujdar. Analyzing Titanic disaster using machine learning algorithms, 2017 International Conference on Computing, Communication and Automation (ICCCA), IEEE, 2017.

[2] L. Eric, and C. Tang, Titanic machine learning from disaster, LamTang-TitanicMachineLearningFromDisaster, 2012.

[3] S. Trevor, Titanic: Getting Started With R-Part 3: Decision Trees, 2014.

[4] S. Bruno, A. David, and B. Torgler, Behavior under extreme conditions: The Titanic disaster, Journal of Economic Perspectives, 25(1), 2011, pp: 209-222.

[5] Information on: https://www.kaggle.com/code/nadintamer/titanic-survival-predictions-beginner

[6] IEEE Reliability Society, International Integrated Reliability Workshop Final Report, Electron Device Society and Reliability Society of the Institute of Electrical and Electronics Engineers, 2002.

[7] C. Thomas, and P. Hart, Nearest neighbor pattern classification, IEEE transactions on information theory, 13(1), 1967, pp: 21-27.

[8] S. Ekin, İlhan Omurca, and Neytullah Acun, A comparative study on machine learning techniques using Titanic dataset, 7th international conference on advanced technologies, 2018.

[9] Information on: https://www.kaggle.com/code/nadintamer/titanic-survival-predictions-beginner

[10] Information on: https://www.kaggle.com/code/nadintamer/titanic-survival-predictions-beginner