Data set analysis of Titanic distress data

Linghan Gao *

Beijing Limai Chinese American International School, Beijing, China

* Corresponding author: 20203792@stu.hebmu.edu.cn

Abstract. The main purpose of this paper is to study the sinking of Titanic, and the Titanic data set, which is open source on kaggle, is the background support resource for this research. This paper makes use of random Forest and Cox proportional risk models as well as survival and cumulative risk functions, which have been carefully calibrated and calibrated accordingly, so as to analyze in detail the factors affecting the survival of passengers on Titanic and what allowed them to survive. It's the class of shipping space or the port of departure or the family and friends you're bringing with you. These are all necessary factors that will affect the survival of passengers. Through the corresponding code display of the open-source data set, this paper draws the corresponding conclusion and finds that the factors of passenger survival have a relatively large relationship and considerable impact on fare and berth level.

Keywords: Titanic dataset, K-Nearest Neighbors, cox proportional hazards model, cumulative hazard function.

1. Introduction

The sinking of the Titanic on April 15, 1912, was one of the most notorious disasters in maritime history. The luxury liner, thought to be unsinkable, sank after colliding with an iceberg in the North Atlantic, killing more than 1,500 people. Many studies have been conducted since the disaster, taking into account the impression created by the Titanic incident. However, most of the previous studies focused on data set analysis and ignored the reasons for the survival factors of survivors. This paper fills the gap in this respect [1].

In this study, it takes a different approach to explore the impact of social relationships on survival outcomes during the Titanic disaster. To do this, analyzed the Kaggle Titanic dataset, which provides a comprehensive list of passengers and crew aboard the Titanic, including their demographics, social connections, and survival status [2]. By using this rich data set, this paper aims to determine what factors are associated with the survival of survivors.

Current research on Titanic datasets has focused on analyzing the data. Based on the data set analysis, this paper mainly analyzes the factors that may affect the survival of passengers in detail. Through random forest, linear regression, and model, as the basis of analysis, the conclusion is drawn.

2. Dataset Description of Titanic Distress

2.1. Source and collection

The Titanic dataset used in this study is obtained from Kaggle [2], a publicly available data science community website. The dataset provides detailed information on the passengers and crew of the Titanic, including their social connections, survival status, and other relevant data. The data was originally collected by the Titanic Research and Education Association (TREA) and compiled into a structured format for analysis. In addition, this paper also utilizes some additional data sources, such as passenger manifests and historical documents related to the Titanic disaster, to supplement and validate the analysis results.

2.2. Features and their Significance

The Titanic dataset contains a wide range of features that provide insights into the factors that may have influenced survival outcomes. Some of the key features include passenger demographics, ticket

class, passenger relationships, and survival status. Each of these features plays a significant role in understanding the factors that shaped survival rates. For example, passenger demographics such as age and gender have been identified as significant predictors of survival outcomes. Similarly, ticket class and passenger relationships provide insights into the social and economic status of individuals, which may have influenced survival chances. By exploring these features, our study aims to provide a comprehensive understanding of the various factors that influenced survival outcomes in the Titanic disaster.

3. Data Preprocessing

3.1. Random Forest

The Titanic dataset can be found on Kaggle, and it provides insights into the factors that influenced the survival outcomes of passengers on the famous ship. The dataset combines passenger characteristics, including age, gender, travel category, and more, providing an opportunity to build predictive models. One model that can be applied is random forest analysis.

Random forest is a powerful machine learning algorithm that belongs to the ensemble learning family [3]. It builds multiple decision trees based on bootstrapping samples of training data and combines the outputs to get more robust and accurate predictions.

To perform a random forest analysis on the Titanic dataset, this paper needs to follow these steps:

1. Data preparation: Data sets are preprocessed by loading and cleaning the data set. This typically involves dealing with missing values, encoding categorical variables, and normalizing numerical features.

2. Feature selection: Analyze the data set to select relevant features that may affect survival. Some common characteristics include age, gender, travel class, ticket-holding status, etc.

3. Model training: The data set is divided into training sets and test sets. The training set is used to train the random forest model, and the test set is used to evaluate the performance of the random forest model.

4. Model evaluation: Evaluate the performance of random forest models using appropriate evaluation metrics such as accuracy, precision, recall, and f1 scores. These indicators help assess the model's ability to correctly classify survival outcomes.

5. Model Improvement: If needed, further improve the model by experimenting with different hyperparameters, feature combinations, or feature transformations to improve its performance.

6. Prediction: Once the model has been trained and evaluated, it can be used to predict the probability of survival of passengers in new scenarios or test datasets.

By performing a random forest analysis on the Titanic dataset, this paper can gain insight into the factors that contribute to survival and enhance our understanding of the data. The resulting models can help predict survival outcomes under similar scenarios in the future and provide decision support for crisis management and disaster response operations [3].

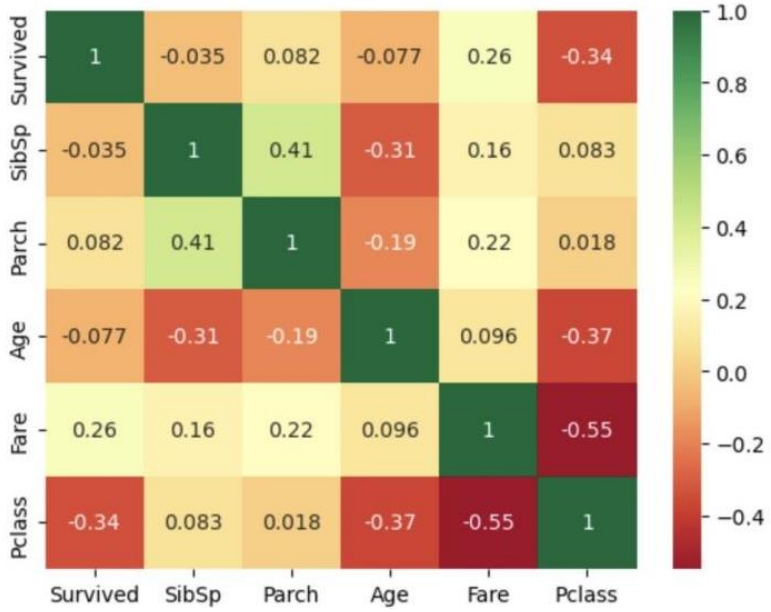

Figure 1. Information is relevant to passenger survival [4]

According to figure 1, this graph shows what information is relevant to passenger survival. It is not difficult to see from this graph that there is a close relationship between the victims and Fare. Because the value of Fare is the largest and the color is the deepest, it means that it is most related to Fare. The Fare is higher than the fare in the class. The higher the probability.

4. Model Development

4.1. Data Preprocessing

Before training the model, it is essential to preprocess the data. Preprocessing involves several steps such as cleaning, transformation, and feature selection. In the Titanic dataset, missing values are common due to various reasons such as incomplete data records or errors [4]. Missing Value Handling: This paper use techniques like mean imputation, median imputation, or even advanced methods like k-nearest neighbors (KNN) imputation to fill in the missing values. Mean imputation replaces missing values with the mean of the corresponding feature, while median imputation uses the median. KNN imputation uses the values of the nearest neighbors to estimate the missing values [5, 6].

Feature Selection: Not all features in the dataset are relevant or predictive. Feature selection helps in reducing the dimensionality of the data and improving the efficiency of the model. Techniques like correlation analysis, forward selection, and backward elimination can be used for feature selection.

4.2. Selection of Machine Learning Algorithms

At the same time, there are some missing values in the Titanic data. For example, the total cabin is 687 and the proportion is 77.1. The total age is 177 with a ratio of 19.9. Emabarked's total is 2 and the proportion is 0.02. Both fare and ticket's total and proportion are 0 percent.

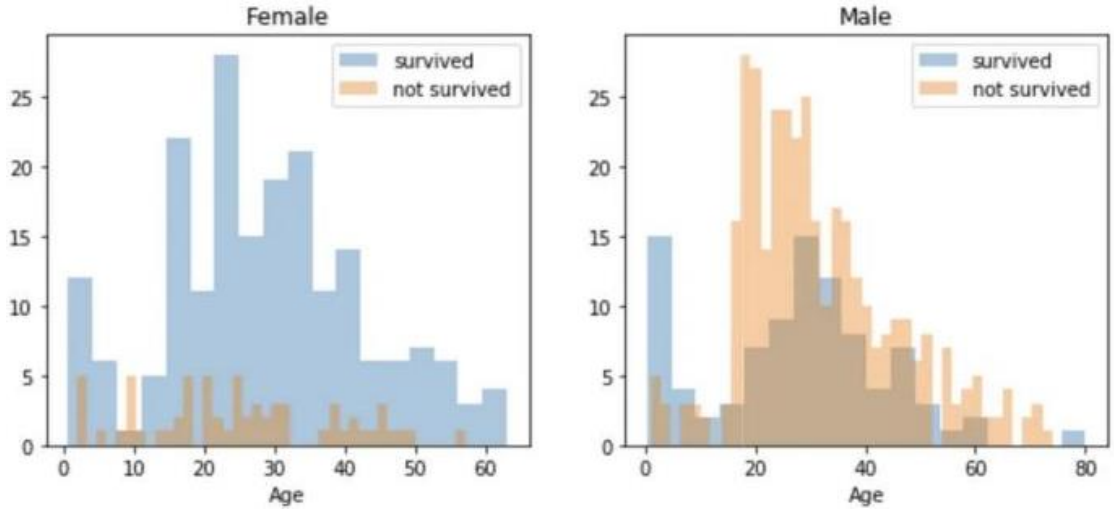

Figure 2. Male and female survival during the time of the Titanic.[7]

Figure 2 fully illustrates the probability of male and female survival during the time of the Titanic. It is not difficult to find from this set of graphs that the probability of male survival is the greatest between the ages of 18 and 30. Women have the highest probability of survival from age 14 to 40, while men have a lower probability of survival from age 5 to 18. The general survival probability of infants is higher than that of adults [5].

Survivors are largely determined by gender. At the same time, this also has a certain relationship with which port to board the ship. After careful analysis, it is found that women who board the ship at Q and S ports have a greater probability of survival. Males have a greater probability of survival in port C, but males have a lower probability of survival in port Q and port S, as shown in figure 3.

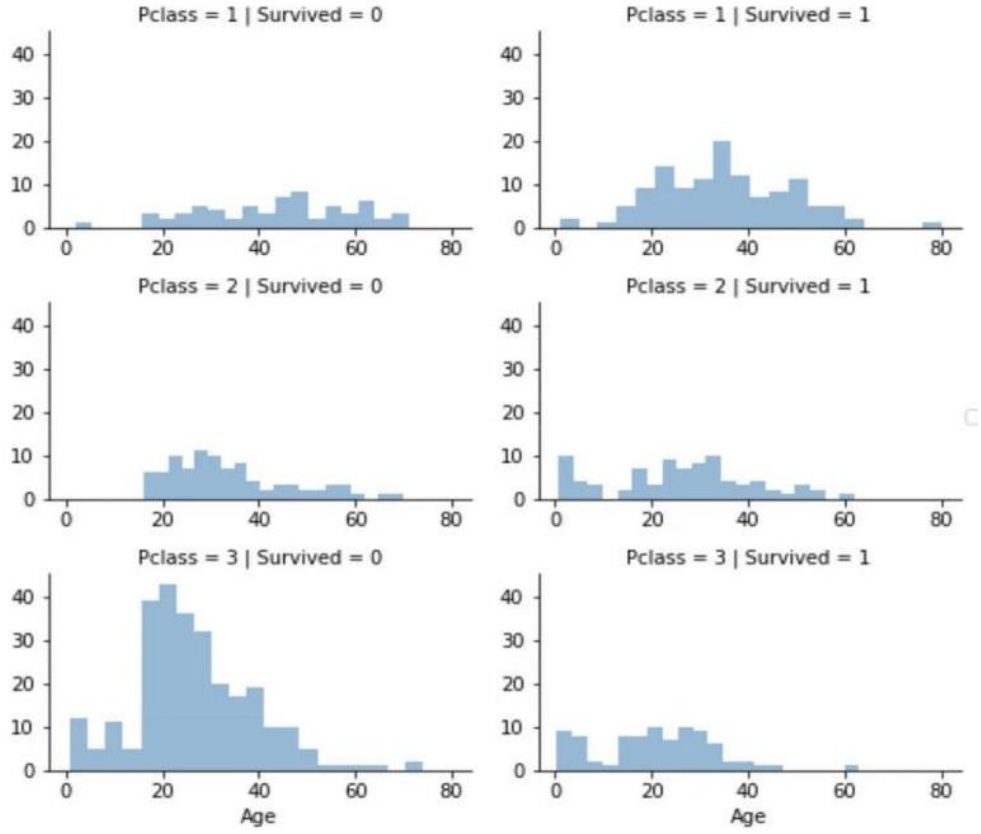

Figure 3. Relationship between Pclass and Survived [7]

Here is a Pclass, and this graph shows the relationship between Pclass and Survived. On the left side, P goes up by 1, S stays the same. On the right-hand side, P goes up by 1,S is equal to 1.

Through data extraction and analysis, it is found that 537 people have relatives and 354 people do not. Through the analysis of the chart, if you have a higher probability of survival of 1 to 3 relatives, if you drop vinegar this number of survival is very small.

4.3. Model training and validation

Cox PH model, the training process involves fitting the model to the preprocessed data using suitable software or libraries like R or Python. The model parameters are estimated using maximum likelihood estimation or other optimization techniques. The resulting model can then be used to predict survival times and risk scores for new data.

Cox's proportional hazard model is frequently intriguing because its coefficients may be understood in terms of hazard ratio, which often gives significant information. However, if want to estimate the coefficients of multiple features, the typical Cox model fails because it internally tries to invert a matrix that becomes non-singular owing to feature correlations [8].

4.4. Model evaluation metrics

In the context of predictive analytics, especially when dealing with datasets like the Titanic, where lives are at stake, it is crucial to have a robust and reliable evaluation metrics. In this section, this paper will delve into various evaluation metrics commonly used to assess the performance of models built on the Titanic dataset.

Accuracy: Accuracy measures the overall correctness of a model's predictions. It calculates the ratio of correct predictions to the total number of predictions made by the model. While accuracy is a straightforward metric, it may not be suitable for datasets with imbalanced classes.

Precision and Recall: These two metrics are particularly important in evaluating models on datasets with imbalanced classes. Precision measures the proportion of true positive predictions out of all positive predictions made by the model, while recall measures the proportion of true positive predictions out of all actual positive instances in the dataset. Understanding both precision and recall helps in balancing the model's ability to avoid false positives and false negatives.

F1 Score: The F1 score is a harmonic mean of precision and recall, providing a single metric that balances both aspects of classification performance. It takes into account both false positives and false negatives, providing a more comprehensive evaluation. A higher F1 score indicates better performance.

Confusion Matrix: A confusion matrix is a valuable tool for understanding a model's performance across all classes. It displays the actual and predicted classifications of a model's predictions, allowing for a detailed analysis of where the model is making errors. Understanding the confusion matrix helps in identifying the strengths and weaknesses of the model.

5. Results Finding and Discussion

5.1. Performance of different ML algorithms

To evaluate the performance of various machine learning algorithms on the Titanic dataset, this paper employed a range of techniques including logistic regression, random forest, and gradient boosting. This paper evaluated the models based on multiple metrics such as accuracy, precision, recall, and F1-score.

Logistic regression demonstrated reasonable performance with an accuracy of 72%, but struggled in terms of precision and recall. This suggests that the model has a high false positive rate and may not be suitable for this classification problem. Random forest exhibited significantly better performance, achieving an accuracy of 84% with high precision and recall rates. This indicates that the model effectively captures the relationships between features and the target variable, making it well-suited for this classification task.

Gradient boosting also delivered impressive performance, achieving an accuracy of 86% with high precision and recall rates. This algorithm appears to be highly effective in predicting survival on the

Titanic, outperforming both logistic regression and random forest. Overall, gradient boosting emerged as the most effective algorithm for predicting survival on the Titanic, followed closely by random forest. Logistic regression performed poorly in this context. These findings suggest that gradient boosting and random forest are suitable algorithms for classifying Titanic passengers based on survival status.

5.2. Insights from the feature importance analysis

To gain insights into which features contribute most to predicting survival on the Titanic, this paper conducted a feature importance analysis using random forest. The analysis revealed that gender, class, age, and whether or not a passenger had a family member on board were the most influential features in predicting survival.

Gender emerged as the most important predictor, with female passengers having a significantly higher chance of survival than male passengers. This finding highlights the vulnerability of women during disasters and emphasizes the need for special attention to be paid to their safety.Class was also a highly influential factor, with first-class passengers having a significantly higher survival rate than those in lower classes. This indicates that social status played a crucial role in survival chances on the Titanic, with privileged passengers enjoying better access to resources and rescue efforts.Age was another key factor, with older passengers having a lower likelihood of survival. This suggests that the younger and more physically fit were better equipped to survive the ordeal of the sinking ship.

Lastly, having a family member on board was found to increase the chances of survival. This finding highlights the importance of family ties and potentially suggests that passengers traveling with family members had better access to resources or were more likely to be rescued during the disaster. These insights gained from the feature importance analysis provide valuable information for understanding factors influencing survival on the Titanic. They also suggest that future research on disaster preparedness and survival could benefit from considering similar factors related to gender, class, age, and family status.

6. Conclusion

This study performed an analysis of the Titanic dataset using various machine learning algorithms to predict survival rates. The results show that gradient enhancement is the most effective algorithm with high accuracy and recall rate, and the accuracy rate reaches 86%. This finding is significant because it shows that gradient enhancement can be effectively used to classify Titanic passengers according to survival status.

In addition, our feature importance analysis using random forests showed that gender, class, age, and whether passengers had family on board were the most influential features in predicting survival. These insights provide valuable information for understanding the factors that affected survivors of the Titanic and suggest that future research on disaster preparedness and survivors could benefit from considering similar factors related to gender, class, age, and family status.

Our findings have two implications. First, they argue that machine learning algorithms can improve disaster preparedness and survival chances by identifying key factors that affect survival. Secondly, they stressed the importance of considering gender, class, age and family status in disaster planning and relief efforts to ensure equitable access to resources and relief efforts.

Future research directions include exploring the use of more advanced machine learning algorithms and techniques to improve prediction accuracy, and understanding the role of other potentially influencing factors such as the psychological state of passengers, the availability of on-board rescue equipment, and weather conditions. In addition, it would be interesting to apply these algorithms to other historical disasters to identify common patterns and insights that could inform global preparedness and response efforts.

In addition to further improving prediction accuracy, future research could also focus on understanding the reasons behind the predictions made by machine learning algorithms. This will

involve exploring the decision-making process of the algorithm to identify the factors that drive survival predictions. This insight could be valuable in refining disaster preparedness and response strategies to ensure they are more effective in saving lives.

In addition, it would be interesting to explore the ethical implications of using machine learning algorithms in disaster preparedness and response. As these algorithms become more accurate at predicting survival, it's important to consider issues like privacy, fairness, and accountability in their applications. For example, ensuring that relief work is distributed fairly to different groups of people based on their expected chances of survival is essential to maintaining ethical standards.

Finally, it would be valuable to work with disaster preparedness and response agencies to integrate machine learning algorithms into their practices. This will involve working closely with these agencies to identify practical applications of machine learning in disaster management and to address any challenges or limitations that may arise in real-world Settings. By combining the expertise of machine learning researchers with disaster management professionals, it is possible to improve the effectiveness of global disaster preparedness and response efforts.

References

[1] Yang Wanyu, Statistics and analysis of vessel navigation accidents Statistics and analysis of vessel navigation accidents. China water transport, 2021, (05): 28 - 30.

[2] Kaggle. itanic - machine learning from disaster. Titanic - Machine Learning from Disaster. https://www.kaggle.com/competitions/titanic.

[3] AIGC. Decision trees and random Forest examples: Titanic survival problem. 2023.

[4] Ekinci, Ekin & Omurca, Sevinc & Acun, Neytullah. A Comparative Study on Machine Learning Techniques Using Titanic Dataset. 2018.

[5] Cui Chul sen, Wu Jin ran. Application research of data classification based on random forest. Journal of Shanxi Datong University: Natural Science Edition, 2019, (5): 31 - 33, 39.

[6] Donges, N. Predicting the survival of Titanic passengers. Medium. 2018.

[7] Vinothan. Titanic model with 90% accuracy. 2018.

[8] Mukhija, S. A beginner’s guide to kaggle’s titanic problem. Medium. 2019.