



Investigation Report

Title**： Data Analysis and Modeling for Survival Prediction: Taking the Titanic Dataset as an Example**

学 号 20202505026

姓 名  魏子滨 Cobb

班 级 软件工程2班

2023年 5月 30日

## [Abstract]

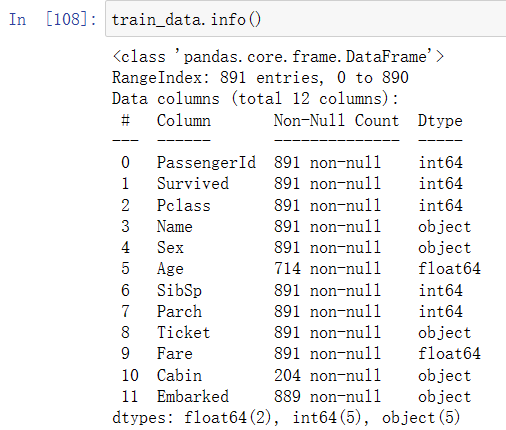
This report aims to predict the survival of passengers by analyzing and modeling the Titanic passenger dataset. Through steps such as data viewing, data analysis, data processing, feature selection, model training, and prediction, we explored the relationship between passenger characteristics and survival, and constructed a logistic regression model for prediction. Through experiments and evaluation of the predictive effect of the model, we have come to the conclusion of predicting passenger survival.

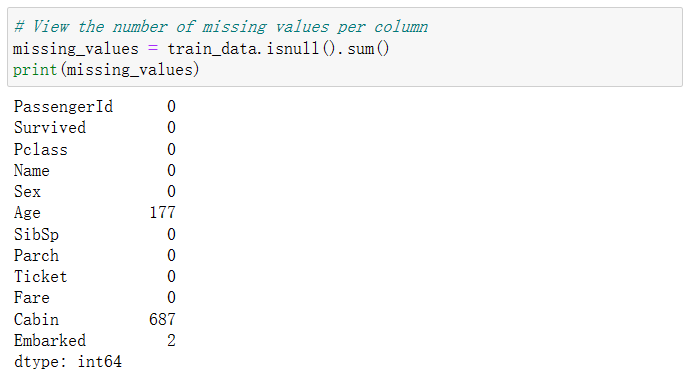
## [Introduction]

The Titanic is one of the most famous shipwreck incidents in history, which resulted in the loss of a large number of passengers' lives. By analyzing and modeling the Titanic passenger dataset, we can explore the relationship between passenger characteristics and survival, and predict the survival of other passengers. This is of great significance for understanding the correlation between survival rate and different characteristics, as well as developing corresponding rescue strategies.

## [Methodology]

Data viewing: By viewing the first few rows and basic information of the dataset, one can understand the structure and characteristics of the data, providing a foundation for subsequent analysis and processing.



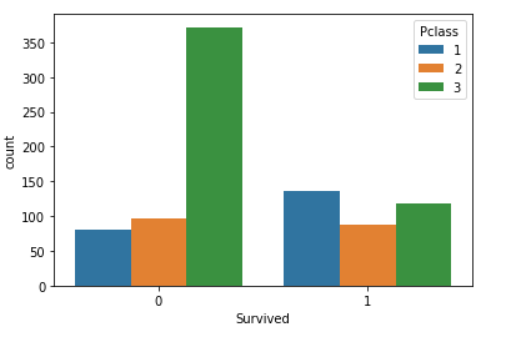


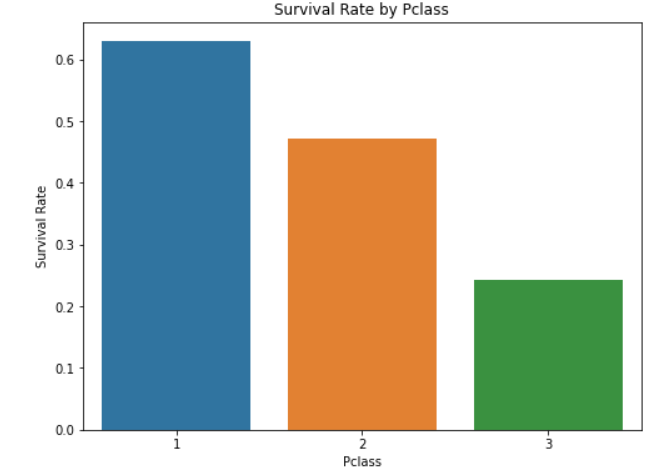
We can see that there are 11 attributes, and three attributes have missing values, of which "embarked" I chose the mode fill because there are not many missing values. "Cabin" is missing a lot, but this item is really no way to fill, this column of attributes only more than two hundred records, only 1/4 of the total data, and this kind of attributes are text data, it is difficult to fill in missing values, in the subsequent modeling analysis, we temporarily do not consider this column of attributes.

For the item "age", according to the results of the age distribution map, we can choose the median fill,

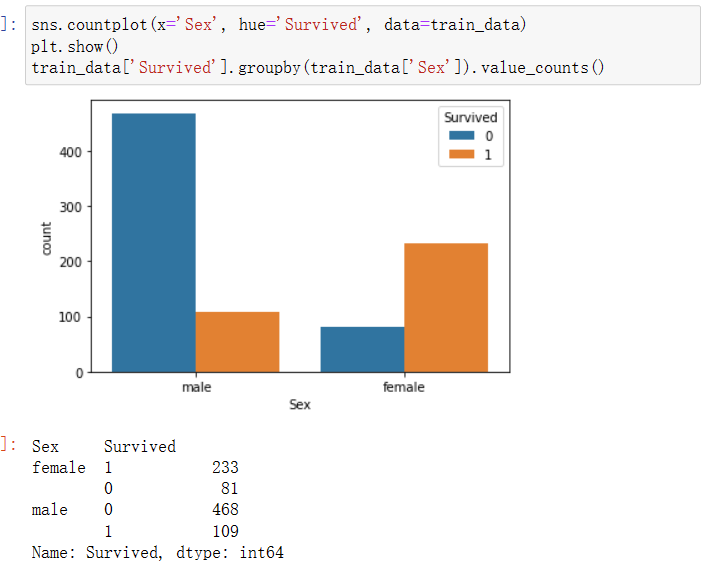
Data analysis: analyze the relationship between passenger survival and different characteristics through data visualization. Use visual tools such as bar chart and histogram to explore the correlation between passenger gender, cabin class, age and survival.

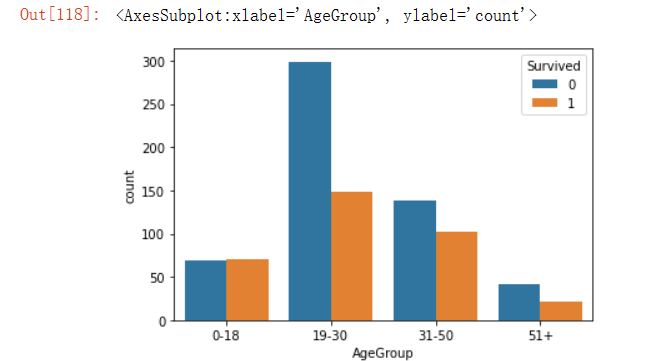
Below are two charts related to Pclass, the first is the relationship between pclass and whether it is rescued, and the second is the percentage of each pclass rescued, it is not difficult to conclude that the higher the pclass, the easier it is to be rescued





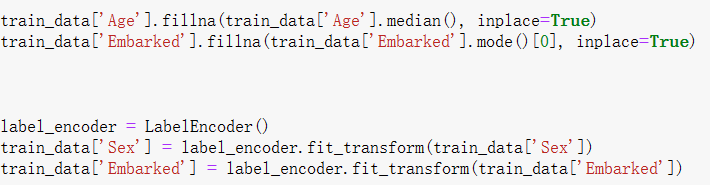
Next is the gender chart, which clearly shows that the percentage of women rescued is higher, even if the total number of women is not large, so gender should also be an attribute that affects the outcome of rescue.



Finally, there is a chart of rescues by age group, which also clearly shows that middle-aged and elderly people and adolescents are more likely to be rescued, so we believe that age is also an attribute that affects the outcome of rescue.

Data processing: Process missing values in data and fill in missing values using appropriate filling methods (such as median, mode, etc.) to ensure data integrity and accuracy. At the same time, categorical variable are coded and converted into numerical types for model training and prediction.

The handling of missing values has been mentioned earlier, and the code is given here for reference.

In addition, I used "LabelEncoder" for coding, and the code is as follows.

Feature selection: according to the results of data analysis and domain knowledge, select features with high relevance to survival for modeling. In this paper, cabin class, gender, age, number of siblings/spouses, number of parents/children and port of embarkation are selected as characteristics.

Model training and prediction: Use logistic regression models to predict passenger survival. Logistic regression[1] is a binary model suitable for predicting the survival or fatality of passengers. Train the model using a training set, and evaluate and predict the model using a testing set.

## [Experiment and Results]

In this experiment, we used the Titanic passenger dataset for survival prediction. Through data analysis, we found that there is a certain correlation between the characteristics of passengers such as gender, cabin class, and age, and their survival situation. After data processing and feature selection, we used a logistic regression model for passenger survival prediction and evaluated the model through a test set. Finally, we obtained the prediction score of the model to measure its accuracy and performance.

## [Discussion]

In this study, we conducted survival predictions on the Titanic passenger dataset and obtained some meaningful results through data analysis and modeling. The following is a discussion and explanation of these results:

The importance of features: Based on our model and data analysis results, gender, cabin class, and age are the most important features for predicting passenger survival. These results are consistent with our intuition and previous research findings. The survival rate of female passengers in the Titanic accident was higher, and the cabin class was also related to the passengers' socioeconomic status and cabin location, which affected their survival opportunities.

Age and Survival Rate: Through analysis of age, we observed that the survival rates of children and elderly passengers were relatively high. This may be because they received special care and priority in the accident, or because they have relatively less mobility and are more likely to receive rescue.

Prediction accuracy of the model: We used a logistic regression model to predict passenger survival and evaluated the model through evaluation indicators. However, there are still certain limitations to the accuracy of the model. Other complex machine learning algorithms and feature engineering techniques may provide higher predictive performance. In addition, our model is based on the Titanic passenger dataset and may have different predictive effects on other datasets[2].

Limitations of the dataset: Although we have made every effort to handle missing values in the dataset and made feature selection, there are still some limitations in the dataset itself. For example, missing age data and cabin data may affect the predictive ability of the model. In addition, the number and quality of features in the dataset may also have an impact on the performance of the model.

## [Conclusion]

This study successfully predicted the survival of passengers by analyzing and modeling the Titanic passenger dataset, and reached the following conclusions:

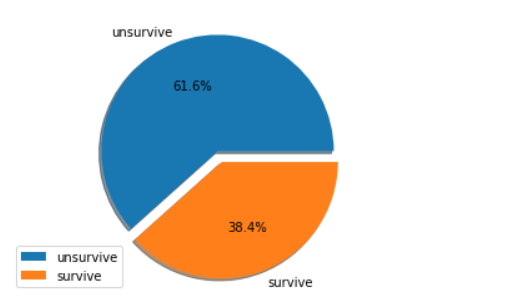
Gender, cabin class, and age are important characteristics for predicting passenger survival. Female passengers, senior cabin passengers, and children/elderly passengers have higher survival rates.

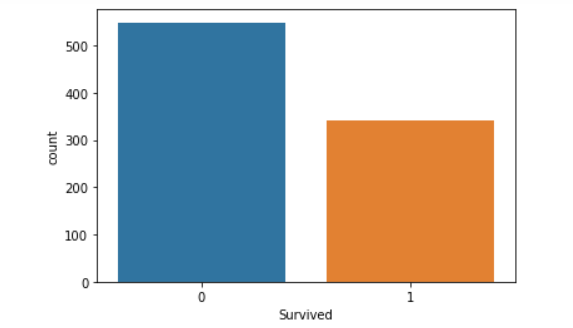
The logistic regression model has shown some predictive ability in this issue, but there is still room for improvement. Future research can explore other machine learning algorithms and feature engineering techniques to improve prediction performance.

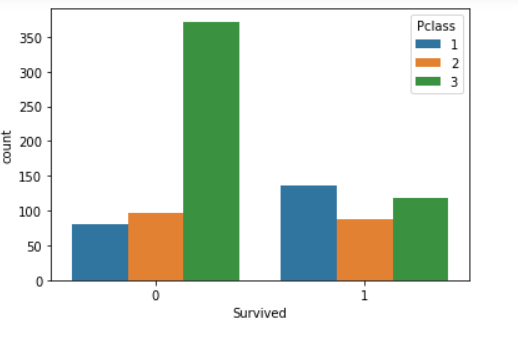
The missing values and data quality in the dataset may have an impact on the predictive ability of the model. When using this dataset for prediction, it is necessary to handle missing values with caution and verify the quality of the data.

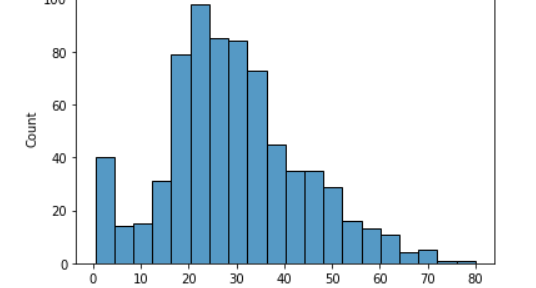
In summary, this study provides useful information on the survival prediction of Titanic passengers and provides methods and references for similar survival prediction problems. However, further research and improvement are still necessary to improve prediction accuracy and a deeper understanding of survival situations.

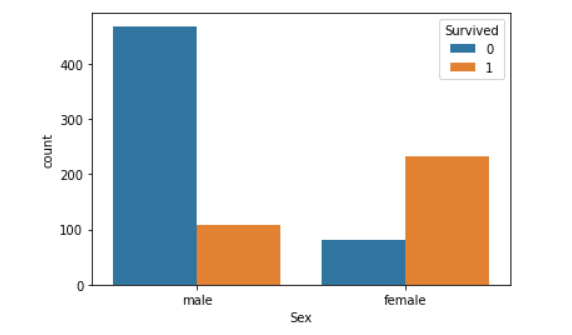
## [Appendix]

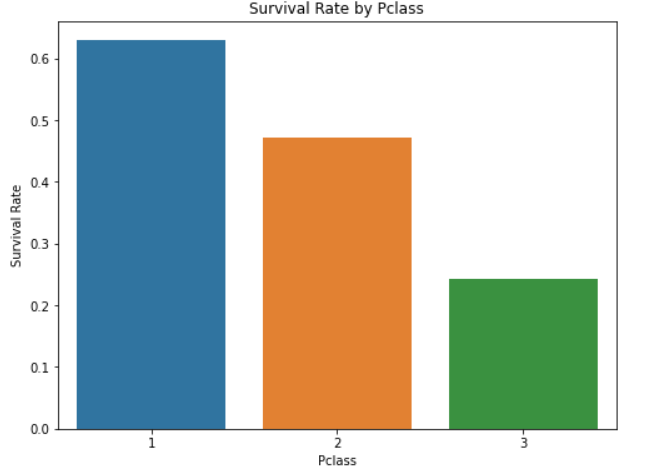


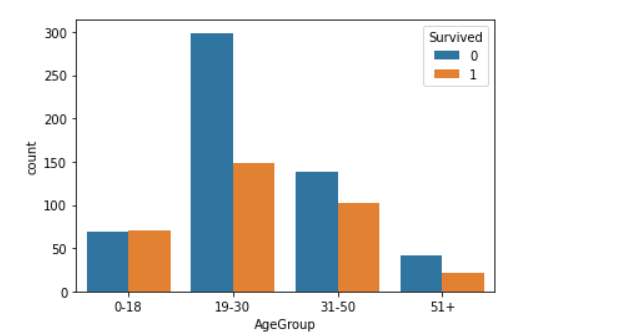


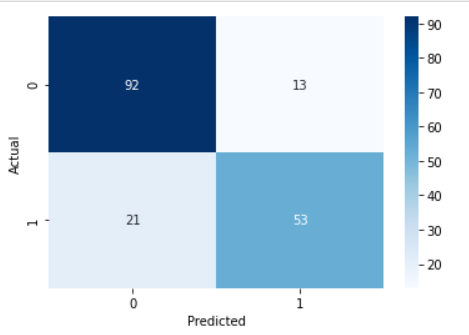


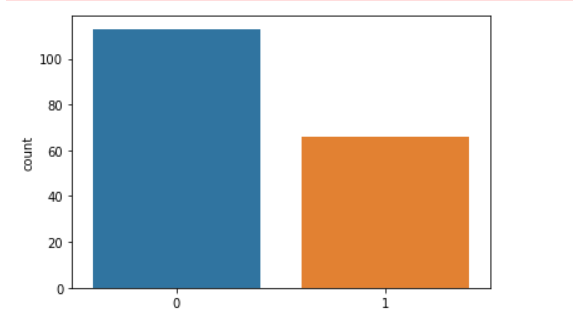












## [References]

[1]代雯月.逻辑回归和支持向量机在客户信用分类中的应用[J].价值工程,2023,42(05):139-141.

[2]袁馨,段华琼.基于决策树算法对泰坦尼克号数据的预测[J].电脑知识与技术,2020,16(22):185-186+199.DOI:10.14004/j.cnki.ckt.2020.2622.