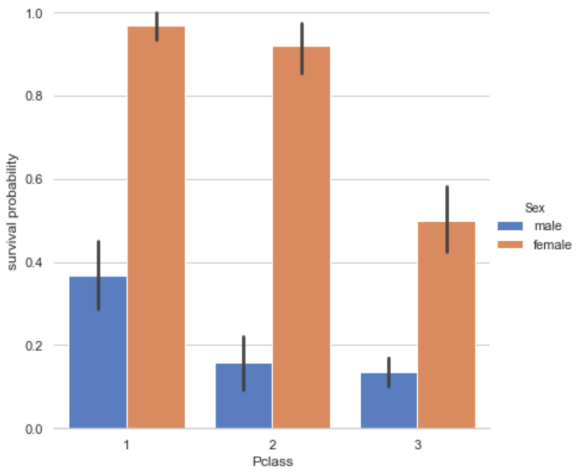
Overview

In the Titanic project, I try to predict the survival result using machine learning model.

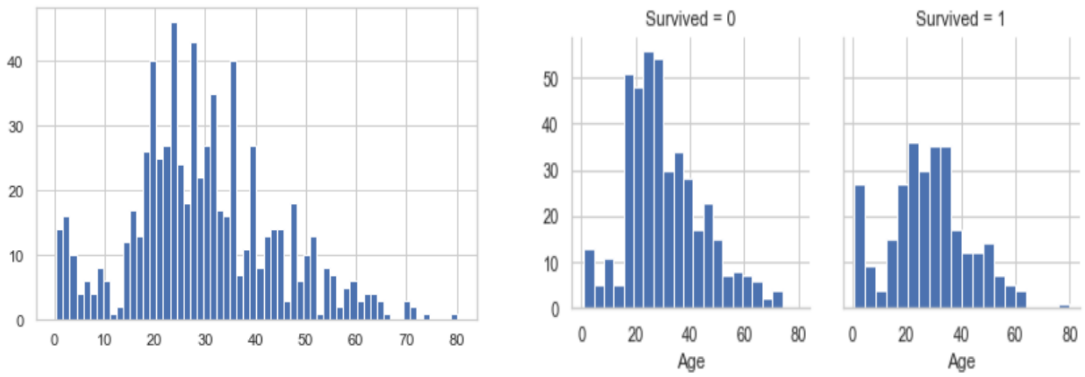
Technique

First, I load data and try to figure out which variables are useful for the upcoming prediction, so I do the analysis for each variable in the chart. For passenger class, I count the number of passengers in each class and calculate the survival rate of each class. The third class has the most passengers, the second class has the least passengers. The survival rate decline as cabin classes goes down.

There are also significant differences in survival rates between the sexes: women are over four times more likely to survive than men. The differences in survival rates between the sexes exist in all three classes. But in the same gender, the survival rate of passengers in the higher-class cabin is higher.



Then, I present the age distribution for all passengers in the sample and draw the age distributions for people dead and survived at the same time. People between the ages of 20 and 40 make up the majority of passengers, and this age group remains the largest group of passengers who died and survived, but one significant change has occurred among passengers under the age of 10: the proportion of passengers under the age of 10 who survived has increased significantly. It is likely that in the face of disaster, more young and middle-aged men have given up the right to be rescued to women and children.



Many passengers board with family members, some with parents, some with siblings. Do passengers with family members have a higher survival rate than those who travel alone? So, I calculated the survival rates for passengers with different numbers of family members. It found that passengers with one sibling had the highest survival rates, which dropped as the number of siblings increased. The survival rates of passengers with parents or children increased and then decreased.

Then I divided the fare interval into four segments and calculated the survival rate for each segment. It found that passenger survival rates increase as fares go up. This may be because the higher the ticket price, the higher the class of the cabin, so there is an increase in the survival rate.

Next is the analysis of the embarkation point. Passengers boarded from three locations: C, Q and S. Most passengers boarded the ship at S. After calculating the survival rates for the three boarding locations, I found that passengers who boarded at C had the highest survival rates, followed by Q and S.

After analyzing each variable, it is the processing of data to prepare for the following model prediction. First, I drop variables that I thought were not highly relevant: Name, Ticket, and Cabin. Then I fill the missing values in Age and Fare with random numbers and average fare. I also convert categorical data into numerical data.

Finally, for the model prediction, I used Decision Tree Classifier to predict passenger survival. I set max depth to 3 and min samples leaf to 2, and finally got the result of this prediction.

Conclusion

My model has reasonable accuracy of 77.99%. The recommendation is to improve the model by changing its parameters or making more improvements in data processing.