1433 Project Report

May 31, 2020

Our project site https://github.com/songrise/COMP1433_GroupProject.

1. Motivation.

The sinking of the Titanic shocked the world, killing 1,514 people and making it the worst peacetime shipwreck in history. It is of great significance and value to analyze the data of this famous shipwreck. I will describe the motivation of analyzing the data of Titanic in the following aspects.

First of all, in terms of history, the sinking of the Titanic is the worst peace wreck. The power of public interest in the immediate aftermath of the Titanic disaster may have had a profound psychological impact, especially in the English-speaking world. Psychologist Wynn Craig Wade said, ″The response to this disaster in the United States was as dramatic as the assassinations of Abraham Lincoln and John F. Kennedy, and shook the entire English-speaking world. At least this tragedy can be seen as a watershed between the 19th and 20th centuries″, literary critic John Wesson foster described the Titanic sinking as ″the end of an era of confidence and optimism that seems to be entering a completely different phase″. According to historian Eric Hobbs, World War I broke out two years after the accident and 19 centuries after human history. A serious shipwreck like the Titanic had a serious impact on society and could even be a potential cause of war. Analyzing the data of Titanic could help to reduce the occurrence of serious shipwrecks.

Besides, after analyzing the data of Titanic, we could figure out the reasons for its occurrence. For instance, disasters have led to major changes in maritime legislation to implement new safety measures, such as ensuring that the provision of life rafts than exercise, lifeboats carrying the correct ship’s permanent passage are equipped with radio equipment by unmanned aerial vehicles. In 1914, to monitor the condition of the icebergs damaging obstacles and the dangerous season of Atlantic maritime security, representatives of 13 countries involved in transatlantic navigation reached a compromise to establish an international iceberg defense service as part of the United States coast, a management fee signed by the United Nations.In the same year, faced with the serious consequences of the Titanic disaster, the consensus reached on maritime safety rules and the international convention on the maintenance of human beings living at sea, two measures remain the global maritime safety norms .

2. Description.

First we process the data and analyse features. There are in total 10 features that may be correlated to survival. Our analysis is based on the assumption that each that each attribute are independently affecting the survival. After finding some of the most relevant attributes, we implemented Naive Bayes model and used the attributes to calculate probability and produced a classification model(a.k.a. NB classifier). Finally, we use the model to predict survival.

The *Naive Bayes* model is based on the Bayes principle and uses the knowledge of probability statistics to classify the sample data sets. It assumes that features are conditionally independent. *Lapalace smoothing* is also adopted to avoid zero probability.

3. Implementation

Instead of simply using others' NB models (E.g. e1071::naiveBayes()), we implemented our own version only with R basic functions. This will be helpful for us to better understand the details of NB classifier.

In the source code file (Scripts/NB.R), there are mainly four parts for our NB implementation, They are:

1. Loading Training Data
2. Calculating Probabilities of Attributes
3. Calculating Conditional Probabilities of Attributes
4. Loading Test Data and Applying NB

Part 1: Loading Training Data

In this section, we used read.csv() and stored the data into a variable called "trainData"

Part 2: Calculating Probabilities of Attributes.

In this part, we calculated the probability of differents attributes. For correlated attributes, we declared some variables to store each of their probabilities, for instance, the probability for Pclass attributes is stored in a list variable, which called pclassProb. pclassProb has three entries corresponding to three possible values of pclass. So that pclassProb[1], pclassProb[2], pclassProb[3] will be p(pclass = 1), p(pclass = 2), p(pclass = 3) respectively.

For some continuous variables, such as age, it is hard and inaccurate to set up probability records for each possible value, so we discreti zed these attributes. For age between 0-10, its probability is ageProb[1] and ageProb[2], ageProb[3] for 10-20, 20-30 ... accordingly.

Another thing to notice is that, for those missing data (NA), we used a technique to set their probability to 1. The idea is that when later applying NB formula (Part 4), multiply a probability of 1 is equivalent to simply ignore that attribute. This technique is implemented by adding extra entry to the probability list variable. For instance, ageProb[9] is set to 1 for NA. To sum up, we use is.na() to identify missing data and set its probability to 1 to ignore those attributes.

Lastly, there is a varible called Lapalace for *Lapalace Smoothing.*

Part 3: Calculating Conditional Probabilities of Attributes

The major calculation part is same as Part2, except that the use of which()function enabled us to extract survived passenger data from trainData.

All variables in this part is named as xxxxProbAlive, for example, pclassProbAlive[1] means p(pclass = 1 | survived).

After this section, The calculation for needed probabilities is done.

Part 4: Loading Test Data and Applying NB

In last section,we used read.csv()to read test data and stored it into a variable called testData, and then start applying NB to predict who will survive using the probabilities calculated in part 2 and part 3. The probability of survival is stored in a variable called p.

We used a loop to extract the attributes of each passenger, the attributes are stored in the variables called pclass, age, sex, etc.

Lastly, we applied the NB formula to calculate the probability of survive for each passenger in testData. If p(survive) < 0.5, he will be considered died, otherwise survived. Then, the prediction result will be exported to a .csv file (Data/prediction.csv) using write.table().

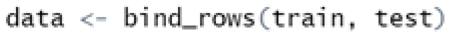
4. Data (Feature analysis)

Load Data

Load the training data and test data into the data.frame named train and test through the following code.

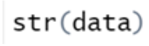
```
train <- read.csv("train.csv")
test <- read.csv("test.csv")
```

In order to avoid repeated operations and inconsistent situations, and to avoid the problem of new levels of Categorical types that may be encountered, it is recommended to merge and unify the training data and test data.

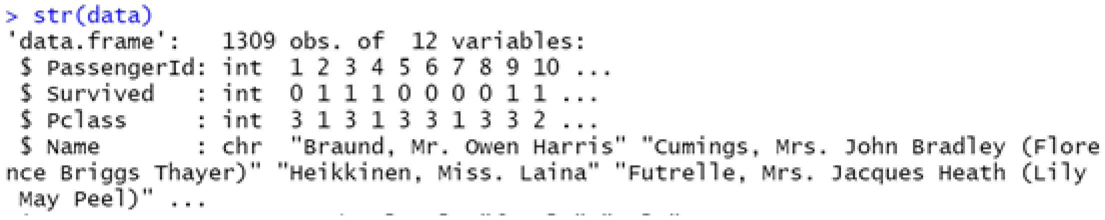


```
data <- bind_rows(train, test)
```

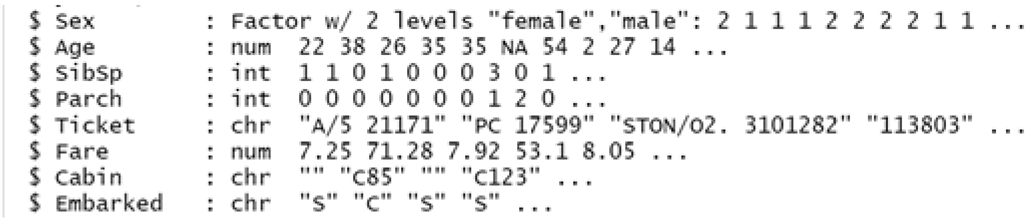
Data preview



```
str(data)
```



```
> str(data)
'data.frame':	1309 obs. of  12 variables:
 $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
 $ Survived   : int  0 1 1 1 0 0 0 0 1 1 ...
 $ Pclass     : int  3 1 3 1 3 3 1 3 3 2 ...
 $ Name       : chr  "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Flore
nce Briggs Thayer)" "Heikkinen, Miss. Laina" "Futrelle, Mrs. Jacques Heath (Lily
 May Peel)" ...
```

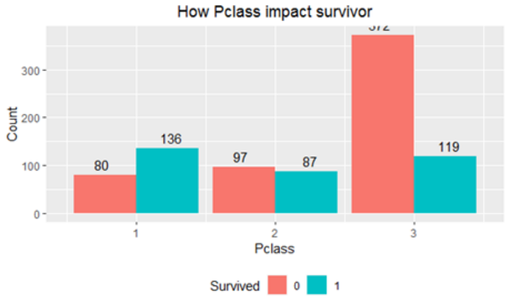


```
 $ Sex        : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
 $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
 $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
 $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
 $ Ticket     : chr  "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...
 $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
 $ Cabin      : chr  "" "C85" "" "C123" ...
 $ Embarked   : chr  "S" "C" "S" "S" ...
```

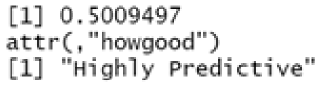
It can be seen from the above that the data set contains 12 variables, 1309 data, of which 891 are training data and 418 are test data

10 variables: **Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, Embarked**

The higher the passenger social rank, the higher the survival rate



It can be seen from the above figure that most passengers with Pclass = 1 survive, nearly half of passengers with Pclass = 2 survive, and less than 25% of passengers with Pclass = 3 survive.

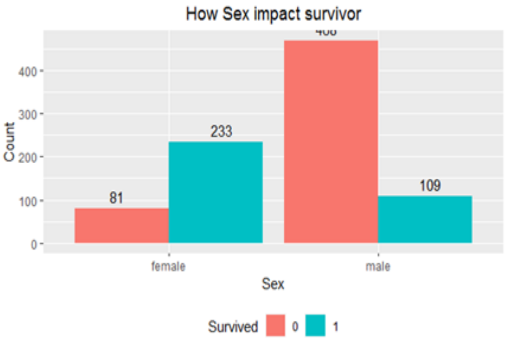


```
[1] 0.5009497
attr(,"howgood")
[1] "Highly Predictive"
```

In order to calculate the predictive value of Pclass more quantitatively, the WOE and IV of Pclass can be calculated. As can be seen from the results, the IV of Pclass is 0.5, and "Highly Predictive". Therefore, Pclass can be temporarily used as one of the characteristic variables of the prediction model.

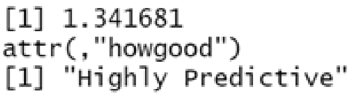
The survival rate of women is much higher than that of men

As for the Sex variable, from the background of the sinking of the Titanic, it can be seen that the rule of ″women and children go first″ is followed when escaping. From this, it is conjectured that the Sex variable should be helpful for predicting passenger survival.



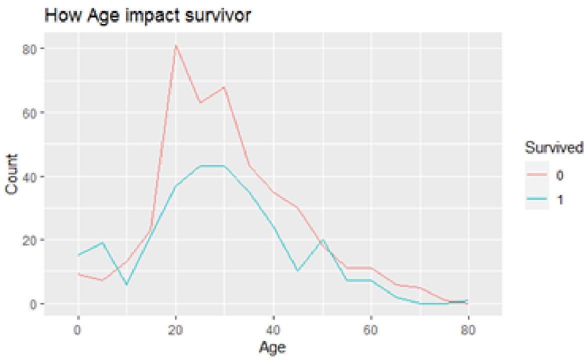
The data validates this conjecture. Most women (233 / (233 + 81) = 74.20%) survived, while only a small percentage of men (109 / (109 + 468) = 22.85%) survived.

By calculating WOE and IV, it can be known that the IV of Sex is 1.34 and ″Highly Predictive″, and Sex can be temporarily used as a characteristic variable.



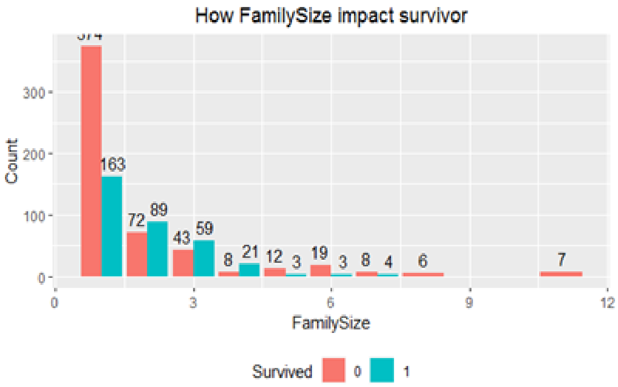
```
[1] 1.341681
attr(,"howgood")
[1] "Highly Predictive"
```

The survival rate of minors is higher than that of adults

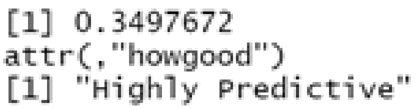


Combined with the background, according to the rule of ″Women and children go first″, minors should be more likely to survive. As shown in the figure below, among passengers with Age <18, the number of survivors is indeed higher than the number of victims. At the same time, among the young and middle-aged passengers, the number of victims was much higher than the number of survivors.

Passengers with FamilySize 2 to 4 are more likely to survive

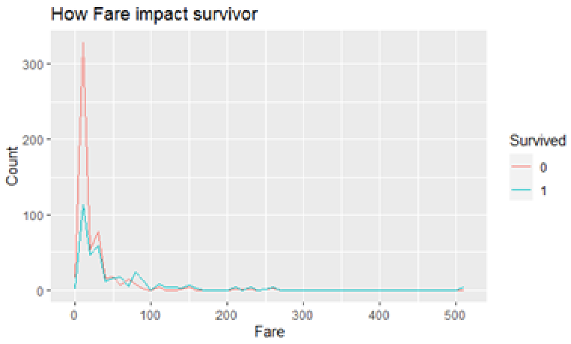


Both SibSp and Parch explained that when passengers have no relatives, the survival rate is lower, when passengers have a few relatives, the survival rate is higher than 50%, and when the number of relatives is too high, the survival rate is reduced. Here, you can consider adding SibSp and Parch to generate a new variable, FamilySize.



```
[1] 0.3497672
attr(,"howgood")
[1] "Highly Predictive"
```

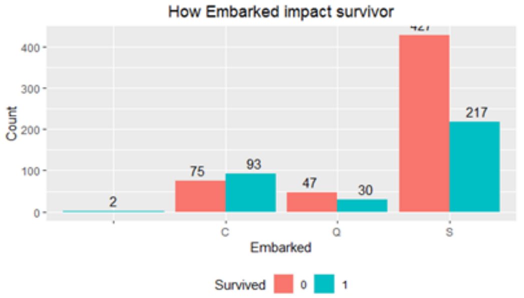
Calculating the WOE and IV of FamilySize, the IV is 0.3497672, and ″Highly Predictive″. The IV of the new variable FamilySize derived from SibSp and Parch is higher than the IV of SibSp and Parch, therefore, this derived variable FamilySize can be used as a characteristic variable.



The higher the fare paid, the higher the survival rate

For the Fare variable, as can be seen from the figure below, the larger the Fare, the higher the survival rate.

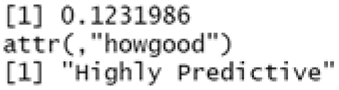
Embarked S passengers have lower survival rates



The Embarked variable represents the embarkation terminal, and it is now possible to determine whether Embarked can be used to predict the survival of passengers by counting the survival rates of passengers boarding at different terminals.

As can be seen from the above figure, the survival rate of passengers with Embarked S is only 217 / (217 + 427) = 33.7%, while the survival rate of passengers with Embarked C or NA is higher than 50%. Preliminary judgments Embarked can be used to predict whether passengers will survive.

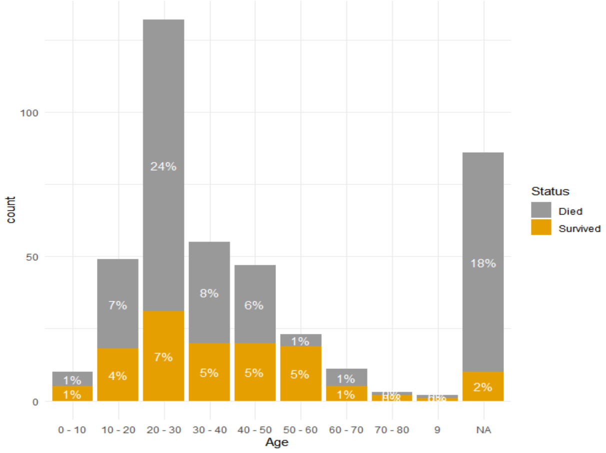
It can be seen from the above calculation results that the IV is 0. 1227284 and ″Highly Predictive″.

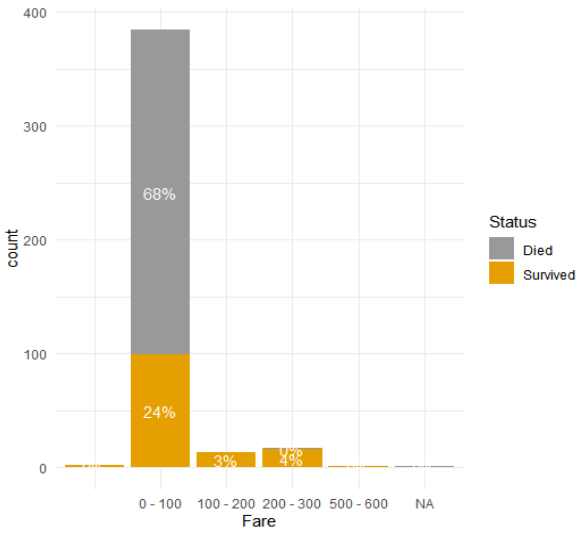


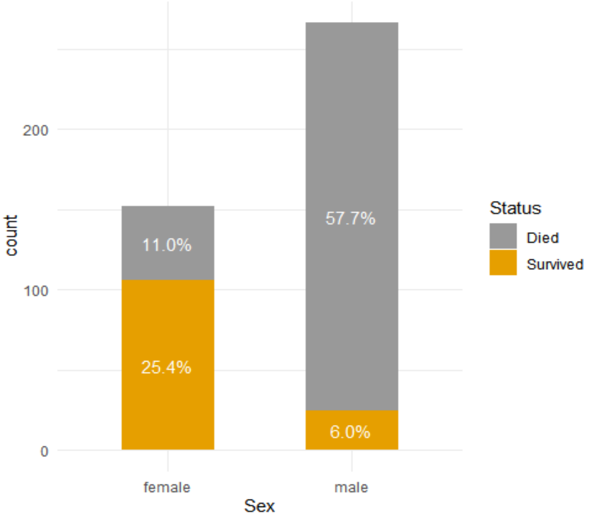
```
[1] 0.1231986
attr(,"howgood")
[1] "Highly Predictive"
```

5. Results and Observations.

We used an R script (resultAnalyze.R) to analyze our prediction. We visualized our prediction so that the relationship between attributes and survived status can be clearly demonstrated. There are mainly three attributes



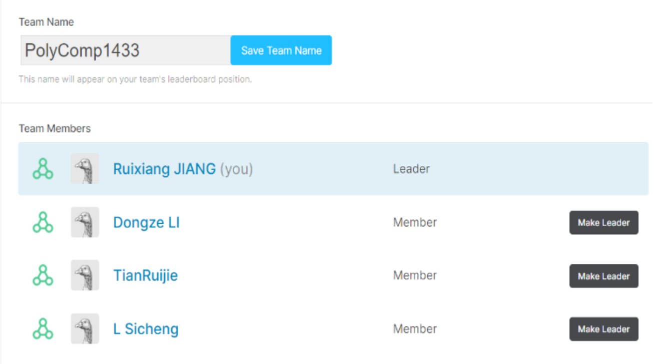




that we visualized, they are: sex, age, fare. *ggplot2* library was adopted to generate these graphs.

As can be clearly seen from the graphs that gender is the dominant attribute for this prediction (This is consistent with former feature analysis). The survival probability of female is significantly higher than that of male.

Moreover, it seems that higher fare will be more likely to survive, but the sample size is still too small to make such assertion.


Team Name

PolyComp1433

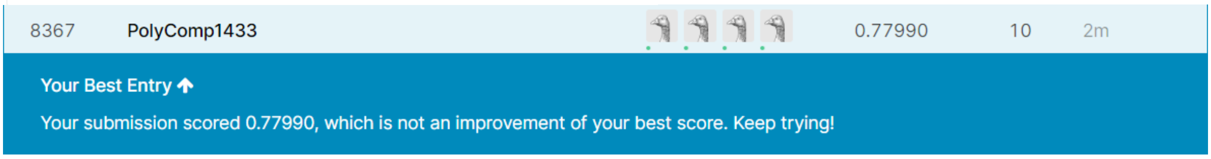
This name will appear on your team's leaderboard position.

Team Members

Ruixiang JIANG (you)	Leader	
Dongze LI	Member	Make Leader
TianRuijie	Member	Make Leader
L Sicheng	Member	Make Leader

Apart from it, unlike most of us would expect, no evident relationship could be found between age and survival, which is quite counterintuitive.

Our latest prediction data submitted to *kaggle.com* got a 0.77990 and ranked 8367 on leaderboard.


8367 PolyComp1433 0.77990 10 2m

Your Best Entry

Your submission scored 0.77990, which is not an improvement of your best score. Keep trying!

6. Discussions.

The disaster was called ″a tragic event that froze time with clockwork determination and became an unforgettable metaphor″ -- the Titanic was more than just a metaphor, it meant so much more; The cultural historian Steven Biel describes it as ″contradictory metaphors, everyone arguing about the social and political significance of the disaster, insisting that it is the real meaning, the real lesson″. The sinking of the Titanic has been explained i n many ways, with some people viewing it in religious terms as a metaphor for divine justice, because what they saw was the greed, arrogance and extravagance of the wealthy passengers. Others interpreted those who stayed on board as "Christian morality that shows self-sacrifice so that women and children can escape". From a social perspective, it can convey information about class or gender relations. There seem to be some "women and children first" principles that affirm chivalrous masculinity and the subordination of women to the "natural" state of men, a view that has been rejected by feminist campaigners. Moreover, the self-sacrifice of multimillionaire John Jacob Astor Iv and Benjamin Guggenheim demonstrated the generosity and moral superiority of the rich and the strong, while the high death toll of third-class passengers and crew became a sign of the low importance of the working class. The main argument is that the behaviour of Anglo-American passengers and crew proves the superiority of "Anglo-Saxon values" in a crisis, the arrogance and hubris of shipowners and Anglo-American elites leading to disaster, or the folly of human beings in their technological progress and excessive complacency. There are so many ways and perspectives to explain this disaster that it remains the subject of public debate and fascination for many years to come.

Apart from a historical view, the experience of analyzing data also enlightened us that the power of ML and Data Analytic are overwhelming. We can see from our prediction that the probability of survive is not randomly distributed. Some people that has certain attributes are more likely to survive in this tragedy (i.e, female, the rich, age between 10-30, etc.) than others. It means that, our analysis may be instructive for future disaster warning and rescue.

7. Teamwork.

Tian Ruijie(19079692d): Process the data and analyse features. (The data part, analyse.R)

LIU Sicheng (19079181d): Read and learn Naïve Bayes. Search and find related documents. Do some optimization.

Ruixiang JIANG (19079662d): Implemented Naïve Bayes classifier (NB.R) and predicted the new data (prediction.csv). Analyzed prediction data by data visualization (resultAnalyze.R). Also managed project files on Github(https://github.com/songrise/COMP1433_GroupProject) for better collaboration.

LI Dongze(19079632d): Write the report (motivation & discussion)