Titanic Survival Rate

David Sierra Perez

2023-07-16

### **Problem**

* What were the highest chances of survival rate on the Titanic based on Passenger Class, Age and Sex?
* Is age also the factor affecting the titanic survival rate?
* What is the possible survival rate of a 20 year old man?

### **What is being tried to accomplish from this project?**

The purpose of this project is to learn the basics of machine learning. We will look at data from a case competition in Kaggle. The competition is using data from the Titanic to predict survival rates during the titanic. We will use different data mining techniques to find patterns in the data.

### **Introduction**

One of the biggest shipwrecks ever recorded was the Titanic. On April 15, 1912, the Titanic that has been said to be “Unsinkable” sank after colliding with an iceberg. To make things worse, the ship was not equipped with enough lifeboats for the passengers and crew members. As a result, over 1500 people died during the incident. Since the Titanic happened in the 1900s. One can assume that women and kids were given priority to get on the lift boats. In this project we will dig deeper into this and try to find out if any other factors played a role in the people who survived. For instance, if passenger class was a factor and age as well.

### **Dataset**

The dataset being used is from Kaggle.com. It contains information from passenger id, if they survived, ticket class, sex, age, number of siblings, number of parents, ticket number, passenger fare, cabin number and lastly port of embark. Kaggle came with 2 data sets. The train data set that tells us whether the passenger survived or died. Survival is given with the value of 1. Did not survive was given with the value of 0. In this project, we will only look at Survived, Pclass (passenger class), Sex, and Age for the the data mining.

### **Literature Survey (existing work on this dataset)**

The Titanic competition has existed for a couple of years now. There have been more than thousands of projects relating to the ship. A lot of work has been done on this dataset. Also, there are currently over 14,000 teams competing in the Kaggle competition. The data we retrieved from Kaggle is more specific and simple. We found out most of the work being done on this dataset has a detailed statistical analysis along with machine learning model implementation. But, the data were more focused on the passenger class and the gender or has related to different predictors like family, cabin number, age and more. However, we limited the predictors to passenger class, gender, and age as we believe that it was an important factor for our project in knowing what age range possibly survived the most. We gathered the data which was pre-processed and specifically analyzed the data based on the passenger class, age and gender.

### **Analysis & Results**

Our first step is to load libraries and upload the data into R.

#install.packages("tidyverse")  
#install.packages("readr")  
#install.packages("ggplot2")  
#install.packages("dplyr")  
#install.packages("janitor")  
#install.packages("skimr")  
#install.packages("here")  
  
library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.2 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ ggplot2 3.4.2 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.2 ✔ tidyr 1.3.0  
## ✔ purrr 1.0.1   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(readr)  
library(ggplot2)  
library(dplyr)  
library(janitor)

##   
## Attaching package: 'janitor'  
##   
## The following objects are masked from 'package:stats':  
##   
## chisq.test, fisher.test

library(skimr)  
library(here)

## here() starts at C:/Users/Sierr/Desktop/My Own Projects/My-Own-Projects

train <- read\_csv("C:/Users/Sierr/Desktop/My Own Projects/My-Own-Projects/Project/Titanic/Dataset/train.csv")

## Rows: 891 Columns: 12  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (5): Name, Sex, Ticket, Cabin, Embarked  
## dbl (7): PassengerId, Survived, Pclass, Age, SibSp, Parch, Fare  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

Lets look at the first 6 rows of data for train dataset

head(train)

## # A tibble: 6 × 12  
## PassengerId Survived Pclass Name Sex Age SibSp Parch Ticket Fare Cabin  
## <dbl> <dbl> <dbl> <chr> <chr> <dbl> <dbl> <dbl> <chr> <dbl> <chr>  
## 1 1 0 3 Braund… male 22 1 0 A/5 2… 7.25 <NA>   
## 2 2 1 1 Cuming… fema… 38 1 0 PC 17… 71.3 C85   
## 3 3 1 3 Heikki… fema… 26 0 0 STON/… 7.92 <NA>   
## 4 4 1 1 Futrel… fema… 35 1 0 113803 53.1 C123   
## 5 5 0 3 Allen,… male 35 0 0 373450 8.05 <NA>   
## 6 6 0 3 Moran,… male NA 0 0 330877 8.46 <NA>   
## # ℹ 1 more variable: Embarked <chr>

Next we will looks at a summary of the data provided

skim\_without\_charts(train)

Data summary

|  |  |
| --- | --- |
| Name | train |
| Number of rows | 891 |
| Number of columns | 12 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 5 |
| numeric | 7 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Name | 0 | 1.00 | 12 | 82 | 0 | 891 | 0 |
| Sex | 0 | 1.00 | 4 | 6 | 0 | 2 | 0 |
| Ticket | 0 | 1.00 | 3 | 18 | 0 | 681 | 0 |
| Cabin | 687 | 0.23 | 1 | 15 | 0 | 147 | 0 |
| Embarked | 2 | 1.00 | 1 | 1 | 0 | 3 | 0 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| PassengerId | 0 | 1.0 | 446.00 | 257.35 | 1.00 | 223.50 | 446.00 | 668.5 | 891.00 |
| Survived | 0 | 1.0 | 0.38 | 0.49 | 0.00 | 0.00 | 0.00 | 1.0 | 1.00 |
| Pclass | 0 | 1.0 | 2.31 | 0.84 | 1.00 | 2.00 | 3.00 | 3.0 | 3.00 |
| Age | 177 | 0.8 | 29.70 | 14.53 | 0.42 | 20.12 | 28.00 | 38.0 | 80.00 |
| SibSp | 0 | 1.0 | 0.52 | 1.10 | 0.00 | 0.00 | 0.00 | 1.0 | 8.00 |
| Parch | 0 | 1.0 | 0.38 | 0.81 | 0.00 | 0.00 | 0.00 | 0.0 | 6.00 |
| Fare | 0 | 1.0 | 32.20 | 49.69 | 0.00 | 7.91 | 14.45 | 31.0 | 512.33 |

From the Summary found using “skim\_without\_charts()” we noticed a couple of things \* There is 891 observations (rows) \* There are 12 columns \* 5 columns are string/character variables \* 7 columns are integers \* we have missing values for Cabin, Embarked and Age.

### **Cleaning Data**

Our analysis is based on the following columns: \* PasssengerId \* Survived \* Pclass \* Age \* Sex

The other columns we will drop.

keep\_columns = c("PassengerId", "Survived", "Pclass", "Age", "Sex")  
  
cleaned\_train <- train[keep\_columns]  
  
skim\_without\_charts(cleaned\_train)

Data summary

|  |  |
| --- | --- |
| Name | cleaned\_train |
| Number of rows | 891 |
| Number of columns | 5 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 1 |
| numeric | 4 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Sex | 0 | 1 | 4 | 6 | 0 | 2 | 0 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| PassengerId | 0 | 1.0 | 446.00 | 257.35 | 1.00 | 223.50 | 446 | 668.5 | 891 |
| Survived | 0 | 1.0 | 0.38 | 0.49 | 0.00 | 0.00 | 0 | 1.0 | 1 |
| Pclass | 0 | 1.0 | 2.31 | 0.84 | 1.00 | 2.00 | 3 | 3.0 | 3 |
| Age | 177 | 0.8 | 29.70 | 14.53 | 0.42 | 20.12 | 28 | 38.0 | 80 |

cleaned\_train <- cleaned\_train %>%   
 drop\_na()

Before we can do further analysis, we need to convert Sex into a boolean variable. To do this i am choosing to name the following genders with the following number. \* Male = 1 \* Female = 0

cleaned\_train$Sex[cleaned\_train$Sex == "male"] <- 1   
cleaned\_train$Sex[cleaned\_train$Sex == "female"] <- 0   
  
head(cleaned\_train)

## # A tibble: 6 × 5  
## PassengerId Survived Pclass Age Sex   
## <dbl> <dbl> <dbl> <dbl> <chr>  
## 1 1 0 3 22 1   
## 2 2 1 1 38 0   
## 3 3 1 3 26 0   
## 4 4 1 1 35 0   
## 5 5 0 3 35 1   
## 6 7 0 1 54 1

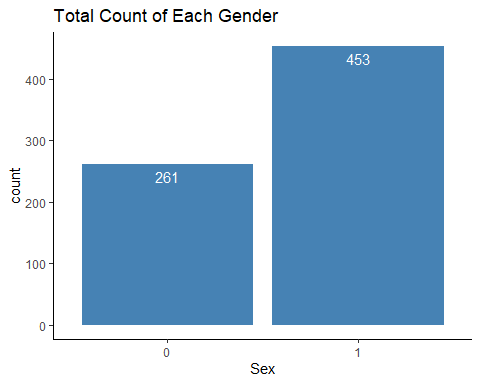
### **Now we will do a quick analysis of our data**

The first step of our analysis was to get a general count of Sex, Passenger Class, Age, and whether someone survived from the train data.

We start of with SEX.

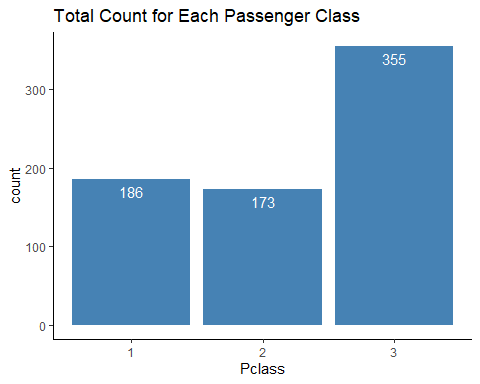
ggplot(cleaned\_train, aes(x = Sex))+   
 geom\_bar(fill = "steelblue") +   
 labs("text", title = "Total Count of Each Gender") +  
 geom\_text(aes(label = ..count..), stat = "count", vjust = 1.5, color = "white") +  
 theme\_classic()

## Warning: The dot-dot notation (`..count..`) was deprecated in ggplot2 3.4.0.  
## ℹ Please use `after\_stat(count)` instead.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was  
## generated.

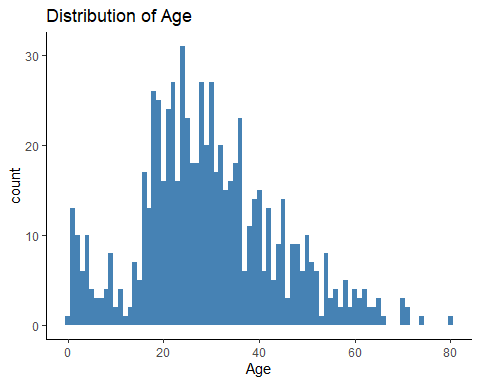


Here is a graph for Passenger Class.

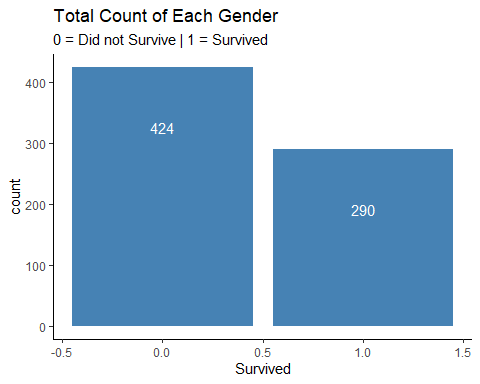
ggplot(cleaned\_train, aes(x = Pclass))+   
 geom\_bar(fill = "steelblue") +   
 labs("text", title = "Total Count for Each Passenger Class") +  
 geom\_text(aes(label = ..count..), stat = "count", vjust = 1.5, color = "white") +  
 theme\_classic()



ggplot(cleaned\_train, aes(x=Age)) +   
 geom\_histogram(binwidth=1, fill = "steelblue" ) +   
 labs("text", title = "Distribution of Age") +  
 theme\_classic()



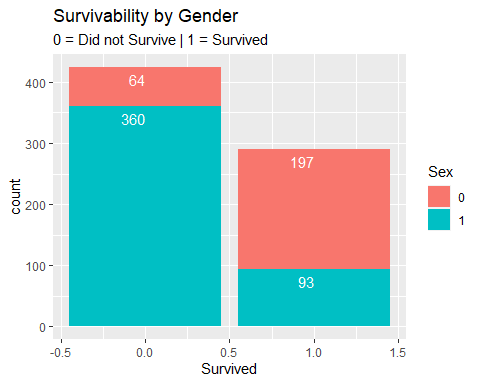
ggplot(cleaned\_train, aes(x = Survived))+   
 geom\_bar(fill = "steelblue") +   
 labs("text", title = "Total Count of Each Gender", subtitle = "0 = Did not Survive | 1 = Survived") +  
 geom\_text(aes(label = ..count..), stat = "count", vjust = 5.5, color = "white") +  
 theme\_classic()



A couple points from the first 4 graphs: \* There are more males than females in the dataset \* There are more passngers in class 3 \* A majority of people are in the ages between 15 to 45 \* More people died than survived

## **Next lets see even more Analysis**

ggplot(cleaned\_train, aes(x = Survived, fill = Sex))+   
 geom\_bar() +   
 geom\_text(aes(label = ..count..), stat = "count", hjust = 1, vjust = 1.5, color = "white", position = "stack")+  
 labs("text", title = "Survivability by Gender", subtitle = "0 = Did not Survive | 1 = Survived")



## **Decision Tree**

We will see try to predict the survival rate of the passenger by Sex, Age, and Passenger Class. Lets take a look at our data again.

cleaned\_train

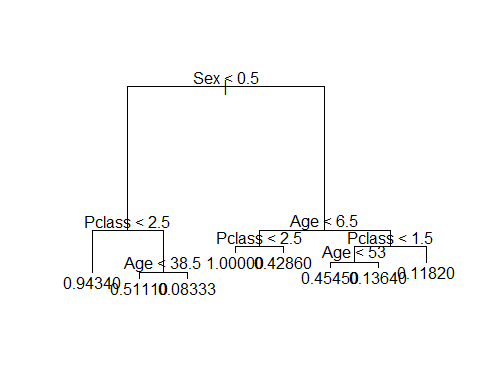
## # A tibble: 714 × 5  
## PassengerId Survived Pclass Age Sex   
## <dbl> <dbl> <dbl> <dbl> <chr>  
## 1 1 0 3 22 1   
## 2 2 1 1 38 0   
## 3 3 1 3 26 0   
## 4 4 1 1 35 0   
## 5 5 0 3 35 1   
## 6 7 0 1 54 1   
## 7 8 0 3 2 1   
## 8 9 1 3 27 0   
## 9 10 1 2 14 0   
## 10 11 1 3 4 0   
## # ℹ 704 more rows

Lets look at a decision tree.

#install.packages("tree")  
library(tree)  
dec\_tree <- tree(Survived ~ Sex + Pclass + Age, data = cleaned\_train)  
  
print(dec\_tree)

## node), split, n, deviance, yval  
## \* denotes terminal node  
##   
## 1) root 714 172.2000 0.40620   
## 2) Sex < 0.5 261 48.3100 0.75480   
## 4) Pclass < 2.5 159 8.4910 0.94340 \*  
## 5) Pclass > 2.5 102 25.3400 0.46080   
## 10) Age < 38.5 90 22.4900 0.51110 \*  
## 11) Age > 38.5 12 0.9167 0.08333 \*  
## 3) Sex > 0.5 453 73.9100 0.20530   
## 6) Age < 6.5 24 5.3330 0.66670   
## 12) Pclass < 2.5 10 0.0000 1.00000 \*  
## 13) Pclass > 2.5 14 3.4290 0.42860 \*  
## 7) Age > 6.5 429 63.1800 0.17950   
## 14) Pclass < 1.5 99 23.4100 0.38380   
## 28) Age < 53 77 19.0900 0.45450 \*  
## 29) Age > 53 22 2.5910 0.13640 \*  
## 15) Pclass > 1.5 330 34.3900 0.11820 \*

plot(dec\_tree)  
text(dec\_tree, pretty = 1)



#female is 0  
#male is 1

logistic\_1 <- glm(Survived ~ Sex + Pclass + Age, family = 'binomial', data = cleaned\_train)  
  
summary(logistic\_1)

##   
## Call:  
## glm(formula = Survived ~ Sex + Pclass + Age, family = "binomial",   
## data = cleaned\_train)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 5.056006 0.502128 10.069 < 2e-16 \*\*\*  
## Sex1 -2.522131 0.207283 -12.168 < 2e-16 \*\*\*  
## Pclass -1.288545 0.139259 -9.253 < 2e-16 \*\*\*  
## Age -0.036929 0.007628 -4.841 1.29e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 964.52 on 713 degrees of freedom  
## Residual deviance: 647.29 on 710 degrees of freedom  
## AIC: 655.29  
##   
## Number of Fisher Scoring iterations: 5

cleaned\_train$Survival\_Rate <- predict(logistic\_1, newdata = cleaned\_train, type = "response")  
cleaned\_train

## # A tibble: 714 × 6  
## PassengerId Survived Pclass Age Sex Survival\_Rate  
## <dbl> <dbl> <dbl> <dbl> <chr> <dbl>  
## 1 1 0 3 22 1 0.105   
## 2 2 1 1 38 0 0.914   
## 3 3 1 3 26 0 0.557   
## 4 4 1 1 35 0 0.922   
## 5 5 0 3 35 1 0.0676  
## 6 7 0 1 54 1 0.321   
## 7 8 0 3 2 1 0.197   
## 8 9 1 3 27 0 0.548   
## 9 10 1 2 14 0 0.877   
## 10 11 1 3 4 0 0.739   
## # ℹ 704 more rows

cleaned\_train <- cleaned\_train %>% mutate(L\_Survived = ifelse(Survival\_Rate > .50, 1,0))  
  
head(cleaned\_train)

## # A tibble: 6 × 7  
## PassengerId Survived Pclass Age Sex Survival\_Rate L\_Survived  
## <dbl> <dbl> <dbl> <dbl> <chr> <dbl> <dbl>  
## 1 1 0 3 22 1 0.105 0  
## 2 2 1 1 38 0 0.914 1  
## 3 3 1 3 26 0 0.557 1  
## 4 4 1 1 35 0 0.922 1  
## 5 5 0 3 35 1 0.0676 0  
## 6 7 0 1 54 1 0.321 0

### **Problem & Answer**

* What were the highest chances of survival rate on the Titanic based on Passenger Class, Age and Sex?
  + Based on our analysis, the survival rate was the highest for the first passenger class, the youngest peple and people who were females.
* Is age also the factor affecting the titanic survival rate?
  + Yes age is a main factor in the titanic. Those who were younger were most likely to survive.
* What is the possible survival rate of a 20 year old man in the first passenger class?
  + Based on our decision tree, the survival rate of a 20 year old man in 1st class is aproximately 45.4%.