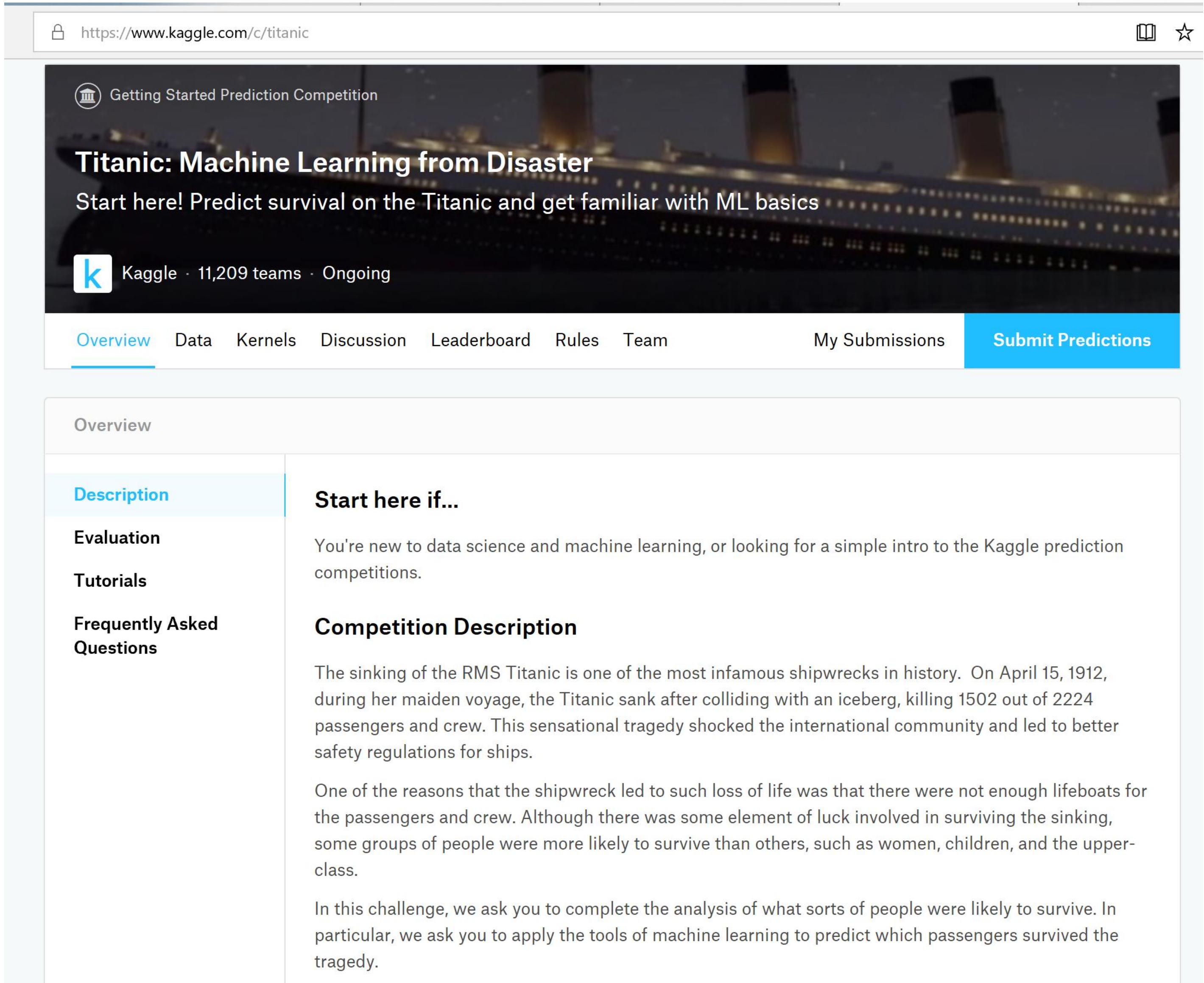
Statistical Machine Learning

Day 10 – Training a Logistic Regression Model

Who survived the Titanic?

An introduction to Kaggle, home of online machine learning competitions

test.csv	418 x 11
train.csv	891 x 12



https://www.kaggle.com/c/titanic

Getting Started Prediction Competition

Titanic: Machine Learning from Disaster

Start here! Predict survival on the Titanic and get familiar with ML basics

Kaggle · 11,209 teams · Ongoing

Overview Data Kernels Discussion Leaderboard Rules Team My Submissions **Submit Predictions**

Overview

Description

Evaluation

Tutorials

Frequently Asked Questions

Start here if...

You're new to data science and machine learning, or looking for a simple intro to the Kaggle prediction competitions.

Competition Description

The sinking of the RMS Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. This sensational tragedy shocked the international community and led to better safety regulations for ships.

One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others, such as women, children, and the upper-class.

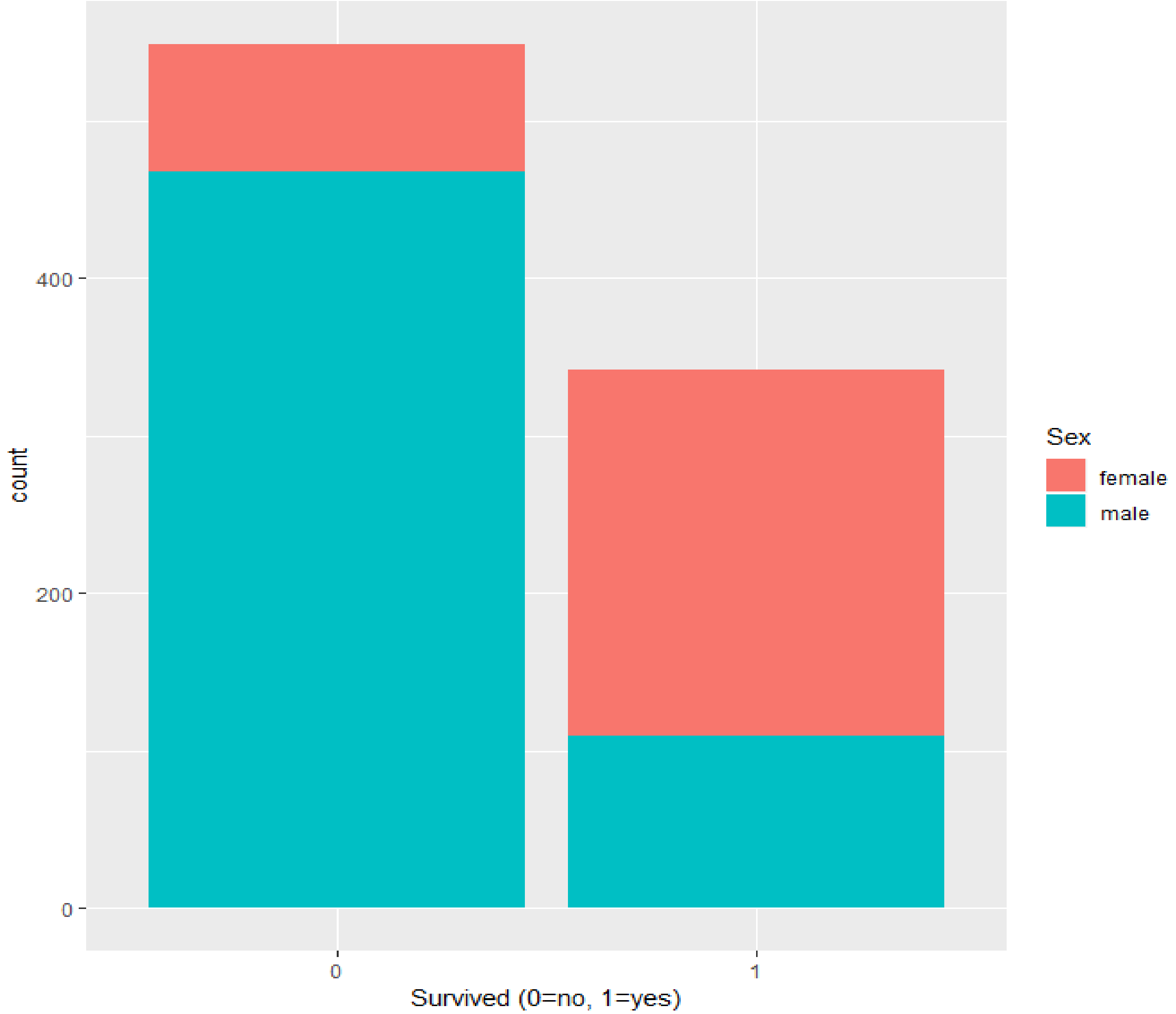
In this challenge, we ask you to complete the analysis of what sorts of people were likely to survive. In particular, we ask you to apply the tools of machine learning to predict which passengers survived the tragedy.

The Training Dataset (R code)

```
> str(dd)
'data.frame':   891 obs. of  13 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       : Factor w/ 891 levels "Abbing, Mr. Anthony",..: 109 191 358 277 16 559 520 629 417 581 ...
 $ 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     : Factor w/ 681 levels "110152","110413",..: 524 597 670 50 473 276 86 396 345 133 ...
 $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
 $ Cabin      : Factor w/ 148 levels "","A10","A14",..: 1 83 1 57 1 1 131 1 1 1 ...
 $ Embarked   : Factor w/ 4 levels "","C","Q","S": 4 2 4 4 4 3 4 4 4 2 ...
 $ Y          : Factor w/ 2 levels "0","1": 1 2 2 2 1 1 1 1 2 2 ...

> table(dd$Survived)/length(dd$Survived)

        0         1
0.6161616 0.3838384
```

What information about a passenger is predictive of their survival?

```
> tab<-table(dd$Survived, dd$Sex)
> tab

    female male
  0     81  468
  1    233  109
>
> tab[2,]/colSums(tab)
   female      male
0.7420382 0.1889081
```



Logistic Regression Model to Predict Survival

```
> fit = glm(Survived~Sex, data=dd, family="binomial")
> summary(fit)

Call:
glm(formula = Survived ~ Sex, family = "binomial", data = dd)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6462  -0.6471  -0.6471   0.7725   1.8256

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.0566     0.1290   8.191 2.58e-16 ***
Sexmale      -2.5137     0.1672 -15.036  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> exp(-2.5137)
[1] 0.0809681
> exp(2.5137)
[1] 12.35054
```

The odds of surviving for a female are estimated to be 12.3 times higher than the odds of survival for a male – we'll see in HW 3 where this interpretation comes from.

How accurate is this model at predicting survival?

```
> probY = predict(fit, type = "response")
> hatY = 1*(probY > 0.5)
> mean(hatY==dd$Survived)
[1] 0.7867565
```

78.7% accurate on the 891 passengers in the training dataset

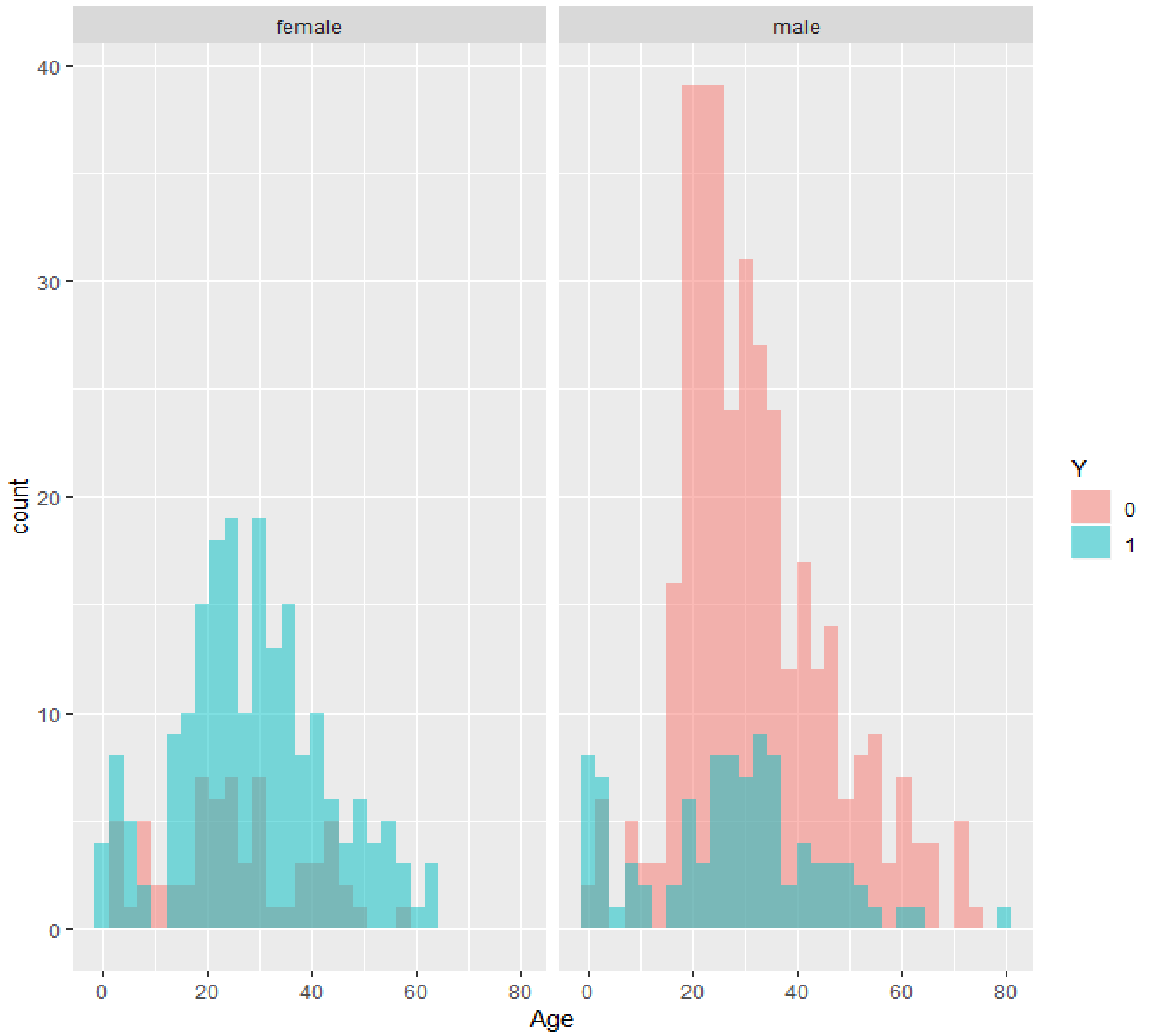
```
> table(hatY,dd$Survived)/rbind(table(dd$Survived), table(dd$Survived))

hatY         0         1
   0 0.8524590 0.3187135
   1 0.1475410 0.6812865
```

85.2% accurate predicting the death of those who didn't survive,
68.1% accurate at predicting the survival of those who did survive.

Who else survived besides females?

Green = Survived

Pink = Didn't Survive



Adding Age to the model…is slightly worse!

```
> fit = glm(Survived~Sex*Age, data=dd, family="binomial")
> summary(fit)

Call:
glm(formula = Survived ~ Sex * Age, family = "binomial", data = dd)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9401  -0.7136  -0.5883   0.7626   2.2455

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.59380    0.31032   1.913  0.05569 .
Sexmale     -1.31775    0.40842  -3.226  0.00125 **
Age          0.01970    0.01057   1.863  0.06240 .
Sexmale:Age -0.04112    0.01355  -3.034  0.00241 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> mean(hatY==dd$Survived[!is.na(dd$Age)])
[1] 0.780112
> table(hatY,dd$Survived[!is.na(dd$Age)])/rbind(table(dd$Survived[!is.na(dd$Age)]), table(dd$Survived[!is.na(dd$Age)]

hatY         0         1
   0 0.8490566 0.3206897
   1 0.1509434 0.6793103
```

78.0% accurate on the 891 passengers in the training dataset

84.9% accurate predicting the death of those who didn't survive, 67.9% accurate at predicting the survival of those who did survive.

Age should be relevant – can we improve the assumed form of the relationship between age and odds of survival?

```
> fit = glm(Survived~Sex*Child, data=ddd, family="binomial")
> summary(fit)

Call:
glm(formula = Survived ~ Sex * Child, family = "binomial", data = ddd)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7244  -0.6226  -0.6226   0.7159   1.8634

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)        1.2305     0.1576   7.806 5.89e-15 ***
Sexmale           -2.7729     0.2031 -13.652  < 2e-16 ***
Childyes          -0.7710     0.4010  -1.923   0.0545 .
Sexmale:Childyes   2.6188     0.5489   4.771 1.83e-06 ***
---
> mean(hatY==ddd$Survived)
[1] 0.7871148
> table(hatY,ddd$Survived)/rbind(table(ddd$Survived), table(ddd$Survived))

hatY         0         1
   0 0.8160377 0.2551724
   1 0.1839623 0.7448276
```