Titanic Data Analysis

Table of contents

```r
library(tidyverse)
```

```
Warning: package 'tidyverse' was built under R version 4.3.3
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.0     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
```

```
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

```r
library(readr)
library(caret)
```

```
Warning: package 'caret' was built under R version 4.3.3

Loading required package: lattice

Attaching package: 'caret'

The following object is masked from 'package:purrr':

    lift
```

```r
library(broom)
```

1 1.

Load Data Convert the Survived,Sex,Cabin and Embarked features to factors

```r
titanic <- read_csv("http://s3.amazonaws.com/notredame.analytics.data/titanic.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.
```

2 2.

Convert the Survived, Sex, Cabin, and Embarked features to factors

```r
titanic <- titanic %>%
  mutate(Survived = as.factor(Survived))

titanic <- titanic %>%
  mutate(Sex = as.factor(Sex))

titanic <- titanic %>%
  mutate(Cabin = as.factor(Cabin))

titanic <- titanic %>%
  mutate(Embarked = as.factor(Embarked))
```

2.1 3.

Which features do you think are useful and which are not? Get rid of any features that are not likely to be useful in the learning process

```r
titanic <- titanic %>%
  select(-PassengerId,-Name,-Ticket,-Cabin,)


summary(titanic)
```

```
Survived     Pclass          Sex           Age            SibSp
0:549    Min.   :1.000   female:314   Min.   : 0.42   Min.   :0.000
1:342    1st Qu.:2.000   male  :577   1st Qu.:20.12   1st Qu.:0.000
         Median :3.000                Median :28.00   Median :0.000
         Mean   :2.309                Mean   :29.70   Mean   :0.523
         3rd Qu.:3.000                3rd Qu.:38.00   3rd Qu.:1.000
         Max.   :3.000                Max.   :80.00   Max.   :8.000
                                      NA's   :177
    Parch             Fare        Embarked
Min.   :0.0000   Min.   :  0.00   C:168
1st Qu.:0.0000   1st Qu.:  7.91   Q: 77
Median :0.0000   Median : 14.45   S:646
Mean   :0.3816   Mean   : 32.20
3rd Qu.:0.0000   3rd Qu.: 31.00
Max.   :6.0000   Max.   :512.33
```

2.2 4.

Are there missing values in the dataset? If so, deal with them appropriately.

```
titanic <- titanic %>%
    group_by(Sex) %>%
    mutate(Age = ifelse(is.na(Age),mean(Age,na.rm = TRUE),Age)) %>%
    ungroup()
```

3 5.

Use a stratified sampling approach to split the dataset into 80% for training and 20% for test.

```
RNGkind(sample.kind = "Rounding")
```

```
Warning in RNGkind(sample.kind = "Rounding"): non-uniform 'Rounding' sampler
used
```

```
set.seed(12334)
sampleset <- createDataPartition(titanic$Survived,p = 0.8,list = FALSE)
titanic_train <- titanic[sampleset,]
titanic_test <- titanic[-sampleset,]
```

4 6.

```
library(performanceEstimation)
```

```
Warning: package 'performanceEstimation' was built under R version 4.3.3
```

```
set.seed(1234)
titanic_train <- smote(Survived ~ .,data = titanic_train,perc.over = 1,perc.under = 2)
titanic_train %>% count(Survived) %>% mutate(prop = round(n/sum(n),4)) %>% arrange(desc(n)
# A tibble: 2 x 3
  Survived     n  prop
  <fct>    <int> <dbl>
1 0          548   0.5
2 1          548   0.5
```

5 7.

Train a logistic regression model using the glm() function from the stats package and display the output.

```
titanic_mod <- glm(Survived ~ .,data = titanic_train,family = binomial)

summary(titanic_mod)
```

```
Call:
glm(formula = Survived ~ ., family = binomial, data = titanic_train)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept)  5.626752   0.566546   9.932  < 2e-16 ***
Pclass      -0.976209   0.143175  -6.818 9.21e-12 ***
Sexmale     -2.535369   0.183609 -13.808  < 2e-16 ***
Age         -0.049209   0.007504  -6.557 5.47e-11 ***
SibSp       -0.489969   0.108214  -4.528 5.96e-06 ***
Parch       -0.215417   0.117874  -1.828   0.0676 .  
Fare         0.010991   0.004552   2.415   0.0158 *  
EmbarkedQ   -0.562007   0.338935  -1.658   0.0973 .  
EmbarkedS   -0.530403   0.220412  -2.406   0.0161 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1519.38  on 1095  degrees of freedom
Residual deviance:  980.52  on 1087  degrees of freedom
AIC: 998.52

Number of Fisher Scoring iterations: 6
```

6 8.

Based on the model output, train a second model with only the significant features from the first model and display the output

```r
titanic_mod2 <- glm(Survived ~ . - Parch - Fare, data = titanic_train,family = binomial)


summary(titanic_mod2)
```

```
Call:
glm(formula = Survived ~ . - Parch - Fare, family = binomial,
    data = titanic_train)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  6.404902   0.468537  13.670  < 2e-16 ***
Pclass      -1.218489   0.112655 -10.816  < 2e-16 ***
Sexmale     -2.490314   0.174759 -14.250  < 2e-16 ***
Age         -0.048538   0.007393  -6.566 5.18e-11 ***
SibSp       -0.450809   0.097335  -4.632 3.63e-06 ***
EmbarkedQ   -0.596678   0.336121  -1.775   0.0759 .
EmbarkedS   -0.650702   0.214337  -3.036   0.0024 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1519.38  on 1095  degrees of freedom
Residual deviance:  990.81  on 1089  degrees of freedom
AIC: 1004.8

Number of Fisher Scoring iterations: 5
```

7 9.

Examine the model coefficients for the second model you created. What impact does Age have on the odds of a passenger surviving the shipwreck?

```
tidy(titanic_mod2) %>%
  select(term,estimate) %>%
  filter(term == "Age") %>%
  mutate(odds = exp(estimate))
```

```
# A tibble: 1 x 3
  term  estimate  odds
  <chr>    <dbl> <dbl>
1 Age    -0.0485 0.953
```

8 10.

What about the gender of the passenger? Who was more likely to survive the accident, men or women?

```
tidy(titanic_mod) %>%
  select(term,estimate) %>%
  filter(term == "Sexmale") %>%
  mutate(odds = exp(estimate))
```

```
# A tibble: 1 x 3
  term    estimate   odds
  <chr>      <dbl>  <dbl>
1 Sexmale    -2.54 0.0792
```