Classify

Thiago Pires

2022-10-26

Table of contents

Introduction

This project presents some analyzes to predict if a passenger given some features could survive or not. I will use data of the famous ship Titanic that tragically wrecked in 1912.

Read data

```r
train <- titanic::titanic_train
```

About the attributes

The variables used are:

- Sex: sex
- Age: age
- Pclass: Passenger Class
- Survived: Passenger Survival Indicator

Initial plan for data exploration

Univariate analysis to identify more important features to multiple model. I will use plots and supervised modeling in the analysis.

Manipulation (Actions taken for data cleaning and feature engineering)

```r
train <- train |>
  dplyr::mutate(Survived = factor(Survived, labels = c("no", "yes")),
         Pclass = factor(Pclass, labels = c("1st", "2nd", "3rd")),
         Sex = factor(Sex))
```

Univariate analysis

My three hypothesis about this data are:

- Relation between sex and survive
- Relation between pclass and survive
- Relation between age and survive

Ticket class (Pclass)

There were more survivors in first class than in the second and third class.

```r
train |>
    ggplot2::ggplot(ggplot2::aes(Pclass, ..count../sum(..count..), fill = Survived)) +
    ggplot2::geom_bar(position = "fill") +
    ggplot2::scale_y_continuous(labels = scales::percent) +
    ggplot2::theme_light() +
    ggplot2::labs(x = "Ticket class", y = "", fill = "Survived")
```

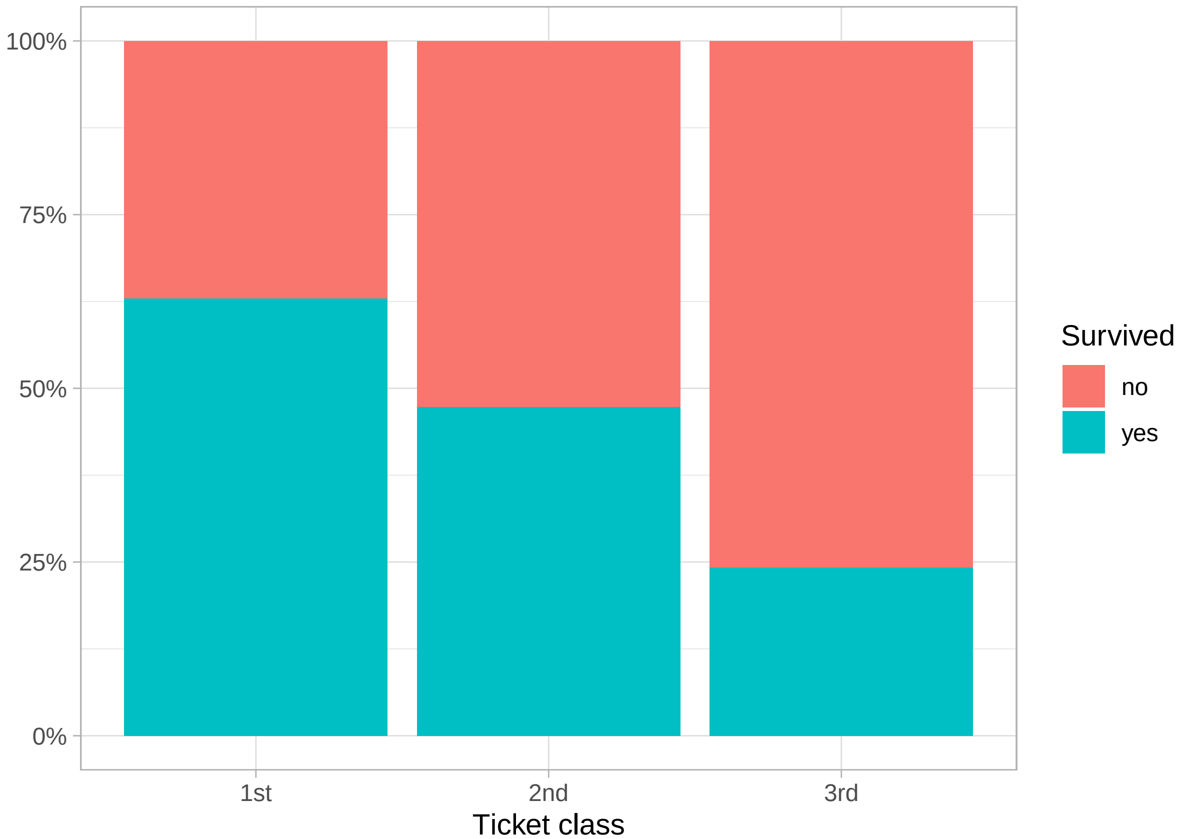

Figure 1: Percentual distribution of survivors according to ticket class

Sex

There were more female survivors than male.

```r
train |>
    ggplot2::ggplot(ggplot2::aes(Sex, ..count../sum(..count..), fill = Survived)) +
    ggplot2::geom_bar(position = "fill") +
```

```
    ggplot2::scale_y_continuous(labels = scales::percent) +
    ggplot2::theme_light() +
    ggplot2::labs(x = "Sex", y = "", fill = "Survived")
```

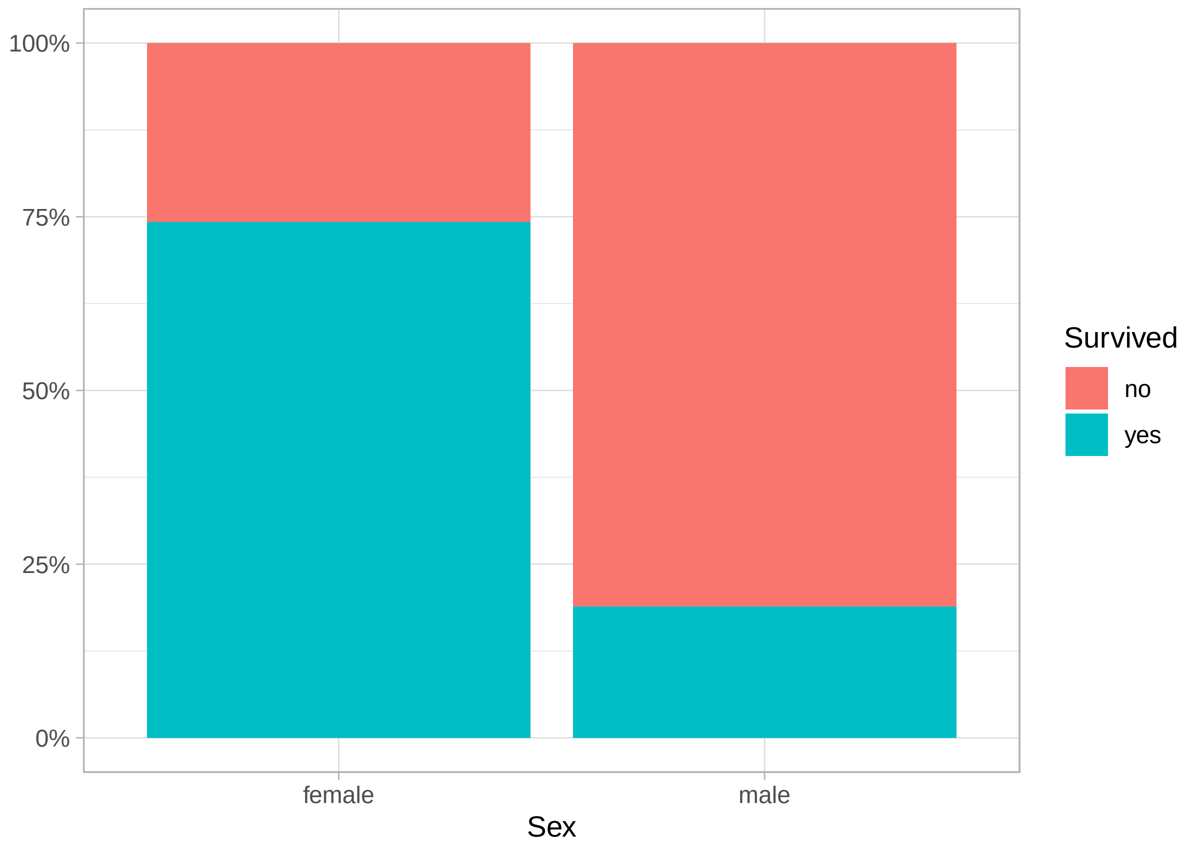

Figure 2: Percentual distribution of survivors according to sex

Interaction between Sex and Ticket Class

There is a diffrence between sex survived and classes. More woman survived in 1st class (about 100%!) than in 3rd class (about 50%).

```
train |>
    ggplot2::ggplot() +
    ggplot2::aes(Sex, ..count../sum(..count..),
        group = Survived,
        fill = Survived) +
```

```
  ggplot2::geom_bar(position = "fill") +
  ggplot2::facet_grid(~Pclass) +
  ggplot2::scale_y_continuous(labels = scales::percent) +
  ggplot2::labs(x = "Sex", y = "", fill = "Survived") +
  ggplot2::theme_light()
```

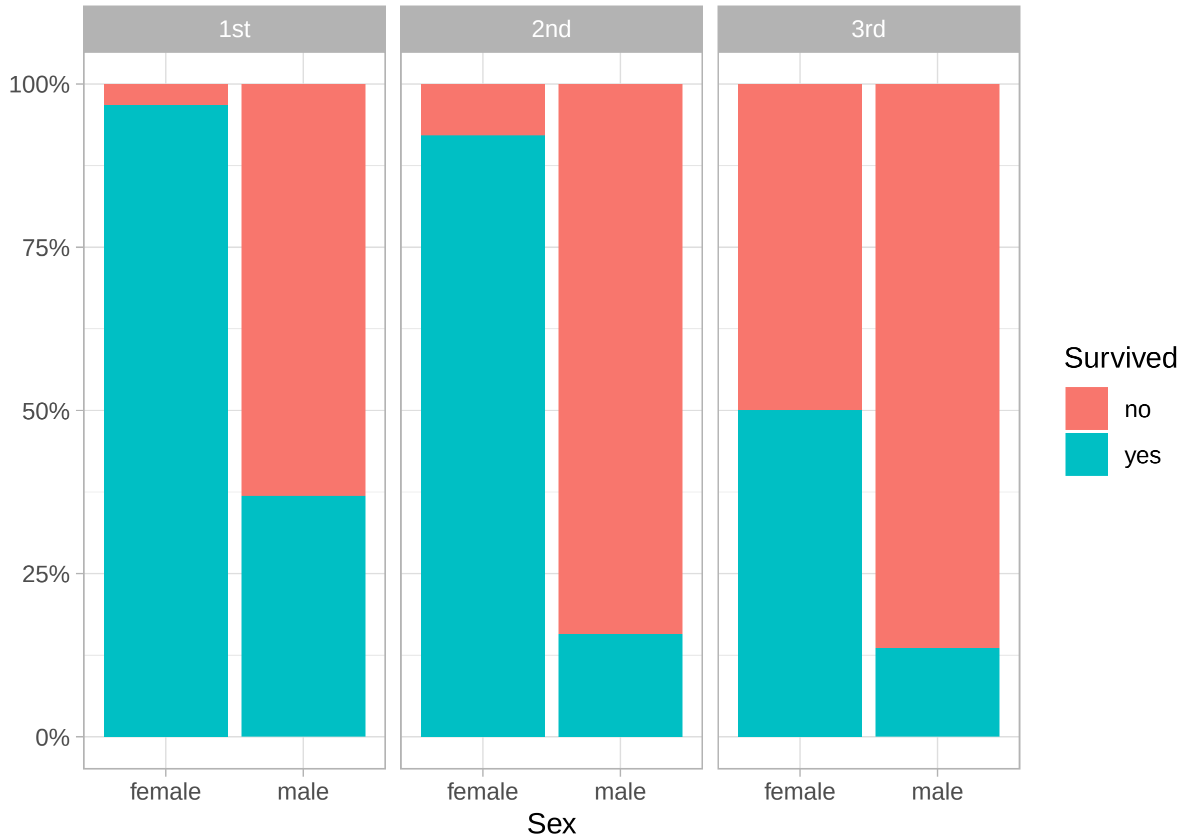

Figure 3: Percentual distribution of survivors according to sex and pclass

Age

There is no difference between the distribution of age according to survived status.

```
train |>
  ggplot2::ggplot(ggplot2::aes(Survived, Age)) + ggplot2::geom_boxplot() +
  ggplot2::theme_light()
```

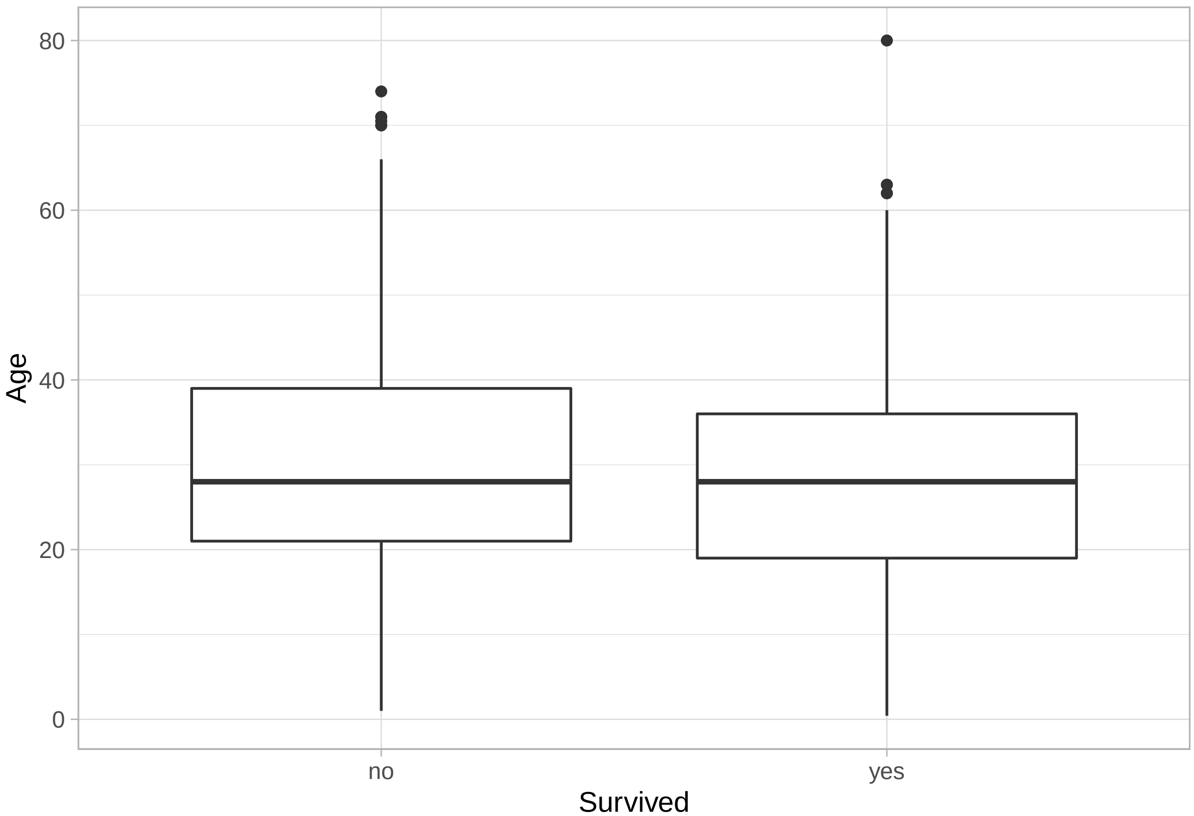

Figure 4: Distribution of age according to survived status

Modeling

Three approaches to modeling:

- Logistic regression
- Logistic regression with interaction effect
- Random forest

Split in train and test

```
library(tidymodels)

data_split <-
    initial_split(train, prop = 3/4)
```

```
train_data <- training(data_split)
test_data  <- testing(data_split)
```

Logistic regression

```
lr_mod <-
    logistic_reg() |>
    set_engine("glm")

lr_fit <-
    lr_mod |>
    fit(Survived ~ Sex + Pclass, data = train_data)

lr_fit |>
    broom::tidy() |> knitr::kable()
```

Table 1: Logistic regression

term	estimate	std.error	statistic	p.value
(Intercept)	2.2934332	0.2550323	8.992716	0.0000000
Sexmale	-2.5613065	0.2115198	-12.109061	0.0000000
Pclass2nd	-0.9280302	0.2790363	-3.325840	0.0008815
Pclass3rd	-1.9923763	0.2499461	-7.971224	0.0000000

Evaluation

```
measure <- function(data) {

    data |>
        accuracy(truth = Survived, .pred_class) |>

        bind_rows(
            data |>
                f_meas(truth = Survived, .pred_class))
}

predict(lr_fit, test_data) |>
    dplyr::bind_cols(predict(lr_fit,
```

```
                        test_data, type = "prob")) |>
    dplyr::bind_cols(test_data |>
    dplyr::select(Survived)) |>
    measure() |>
    knitr::kable()
```

Table 2: Evaluation

.metric	.estimator	.estimate
accuracy	binary	0.8026906
f_meas	binary	0.8394161

Logistic regression with interaction effect

```
lr_fit_i <-
    lr_mod |>
    fit(Survived ~ Sex * Pclass, data = train_data)

lr_fit_i |>
    broom::tidy() |> knitr::kable()
```

Table 3: Logistic regression with interaction effect

term	estimate	std.error	statistic	p.value
(Intercept)	3.4812401	0.7179015	4.8491892	0.0000012
Sexmale	-3.9438636	0.7505417	-5.2546896	0.0000001
Pclass2nd	-1.0833448	0.8564837	-1.2648750	0.2059162
Pclass3rd	-3.6147715	0.7440752	-4.8580730	0.0000012
Sexmale:Pclass2nd	-0.1931474	0.9337726	-0.2068463	0.8361299
Sexmale:Pclass3rd	2.2448135	0.7961158	2.8197073	0.0048067

Evaluation

```
predict(lr_fit_i, test_data) |>
    dplyr::bind_cols(predict(lr_fit_i,
                        test_data, type = "prob")) |>
    dplyr::bind_cols(test_data |>
```

```
    dplyr::select(Survived)) |>
    measure() |>
    knitr::kable()
```

Table 4: Evaluation

.metric	.estimator	.estimate
accuracy	binary	0.7713004
f_meas	binary	0.8370607

Random forest

```
rf <-
    rand_forest(mode = "classification", mtry = 2, trees = 100) |>
    fit(Survived ~ Sex + Pclass, data = train_data)
```

Evaluation

```
predict(rf, test_data) |>
    dplyr::bind_cols(predict(rf,
                    test_data, type = "prob")) |>
    dplyr::bind_cols(test_data |>
    dplyr::select(Survived)) |>
    measure() |>
    knitr::kable()
```

Table 5: Evaluation

.metric	.estimator	.estimate
accuracy	binary	0.7713004
f_meas	binary	0.8370607

Comparing models

The three models get close metrics. But the logistic regression with iteraction get a $f1_{score}$ better than the others models. The results showed that the class and sex are the main effects on the target. Women and class (1st) have more effect in survive.