# Titanic Exercises Part 1

### Titanic Exercises

These exercises cover everything you have learned in this course so far. You will use the background information to provided to train a number of different types of models on this dataset.

### Background

The Titanic was a British ocean liner that struck an iceberg and sunk on its maiden voyage in 1912 from the United Kingdom to New York. More than 1,500 of the estimated 2,224 passengers and crew died in the accident, making this one of the largest maritime disasters ever outside of war. The ship carried a wide range of passengers of all ages and both genders, from luxury travelers in first-class to immigrants in the lower classes. However, not all passengers were equally likely to survive the accident. You will use real data about a selection of 891 passengers to predict which passengers survived.

### Libraries and data

Use the **`titanic_train`** data frame from the **`titanic`** library as the starting point for this project.

```
library(titanic)    # loads titanic_train data frame
library(caret)
library(tidyverse)
library(rpart)

# 3 significant digits
options(digits = 3)

# clean the data - `titanic_train` is loaded with the titanic package
titanic_clean <- titanic_train %>%
    mutate(Survived = factor(Survived),
           Embarked = factor(Embarked),
           Age = ifelse(is.na(Age), median(Age, na.rm = TRUE), Age), # NA age to
median age
           FamilySize = SibSp + Parch + 1) %>%    # count family members
    select(Survived,  Sex, Pclass, Age, Fare, SibSp, Parch, FamilySize, Embarked)
```

## Question 1: Training and test sets

3/3 points (graded)

Split `titanic_clean` into test and training sets - after running the setup code, it should have 891 rows and 9 variables.

Set the seed to 42, then use the `caret` package to create a 20% data partition based on the `Survived` column. Assign the 20% partition to `test_set` and the remaining 80% partition to `train_set`.

How many observations are in the training set?

712 ✔ **Answer:** 712

712

**Explanation**

The following code will give the number of observations in the training set:

```
#set.seed(42)
set.seed(42, sample.kind = "Rounding")    # simulate R 3.5
test_index <- createDataPartition(titanic_clean$Survived, times = 1, p = 0.2, list =
FALSE)    # create a 20% test set
test_set <- titanic_clean[test_index,]
train_set <- titanic_clean[-test_index,]

nrow(train_set)
```

How many observations are in the test set?

179 ✔ **Answer:** 179

179

**Explanation**

The following code will give the number of observations in the test set:

```
nrow(test_set)
```

What proportion of individuals in the training set survived?

0.383 ✔ **Answer:** 0.383

0.383

**Explanation**

The following code will give the survival proportion:

```
mean(train_set$Survived == 1)
```

Submit You have used 1 of 10 attempts

Answers are displayed within the problem

## Question 2: Baseline prediction by guessing the outcome

1/1 point (graded)

The simplest prediction method is randomly guessing the outcome without using additional predictors. These methods will help us determine whether our machine learning algorithm performs better than chance. How accurate are two methods of guessing Titanic passenger survival?

Set the seed to 3. For each individual in the test set, randomly guess whether that person survived or not by sampling from the vector `c(0,1)`. Assume that each person has an equal chance of surviving or not surviving.

What is the accuracy of this guessing method?

0.475 ✔ **Answer:** 0.542 **or** 0.475

0.475

**Explanation**

The accuracy can be calculated using the following code:

```
#set.seed(3)
set.seed(3, sample.kind = "Rounding")
# guess with equal probability of survival
guess <- sample(c(0,1), nrow(test_set), replace = TRUE)
mean(guess == test_set$Survived)
```

Submit You have used 1 of 10 attempts

Answers are displayed within the problem

## Question 3a: Predicting survival by sex

2/2 points (graded)

Use the training set to determine whether members of a given sex were more likely to survive or die. Apply this insight to generate survival predictions on the test set.

What proportion of training set females survived?

0.731 ✔ **Answer:** 0.733 **or** 0.731

0.731

**Explanation**
The proportion can be calculated using the following code:

```
train_set %>%
    group_by(Sex) %>%
    summarize(Survived = mean(Survived == 1)) %>%
    filter(Sex == "female") %>%
    pull(Survived)
```

What proportion of training set males survived?

0.197 ✔ **Answer:** 0.193 **or** 0.197

0.197

**Explanation**
The proportion can be calculated using the following code:

```
train_set %>%
    group_by(Sex) %>%
    summarize(Survived = mean(Survived == 1)) %>%
    filter(Sex == "male") %>%
    pull(Survived)
```

Submit You have used 1 of 10 attempts

Answers are displayed within the problem

## Question 3b: Predicting survival by sex

1/1 point (graded)
Use the training set to determine whether members of a given sex were more likely to survive or die. Apply this insight to generate survival predictions on the test set.

Predict survival using sex on the test set: if the survival rate for a sex is over 0.5, predict survival for all individuals of that sex, and predict death if the survival rate for a sex is under 0.5.

What is the accuracy of this sex-based prediction method on the test set?

0.821

✔ **Answer:** 0.81 **or** 0.821

0.821

**Explanation**
The accuracy can be calculated using the following code:

```
sex_model <- ifelse(test_set$Sex == "female", 1, 0)    # predict Survived=1 if female, 0
if male
mean(sex_model == test_set$Survived)    # calculate accuracy
```

Submit You have used 1 of 10 attempts

Answers are displayed within the problem

## Question 4a: Predicting survival by passenger class

1/1 point (graded)
In which class(es) ( `Pclass` ) were passengers more likely to survive than die?
Select ALL that apply.

- [x] 1
- [ ] 2
- [ ] 3

✔

**Explanation**
The survival rates by class can be found using the following code:

```
train_set %>%
    group_by(Pclass) %>%
    summarize(Survived = mean(Survived == 1))
```

Submit You have used 1 of 2 attempts

Answers are displayed within the problem

## Question 4b: Predicting survival by passenger class

1/1 point (graded)
Predict survival using passenger class on the test set: predict survival if the survival rate for a class is over 0.5, otherwise predict death.

What is the accuracy of this class-based prediction method on the test set?

0.704 ✔ **Answer:** 0.682 **or** 0.704

0.704

**Explanation**
The accuracy can be found using the following code:

```
class_model <- ifelse(test_set$Pclass == 1, 1, 0)    # predict survival only if first
class
mean(class_model == test_set$Survived)    # calculate accuracy
```

Submit You have used 1 of 10 attempts

Answers are displayed within the problem

## Question 4c: Predicting survival by passenger class

1/1 point (graded)
Group passengers by both sex and passenger class.

Which sex and class combinations were more likely to survive than die?
Select ALL that apply.

- [x] female 1st class
- [x] female 2nd class
- [ ] female 3rd class
- [ ] male 1st class
- [ ] male 2nd class
- [ ] male 3rd class

✔

**Explanation**
The combinations can be found using the following code:

```
train_set %>%
    group_by(Sex, Pclass) %>%
    summarize(Survived = mean(Survived == 1)) %>%
    filter(Survived > 0.5)
```

Submit You have used 1 of 3 attempts

Answers are displayed within the problem

## Question 4d: Predicting survival by passenger class

1/1 point (graded)
Predict survival using both sex and passenger class on the test set. Predict survival if the survival rate for a sex/class combination is over 0.5, otherwise predict death.

What is the accuracy of this sex- and class-based prediction method on the test set?

0.821 ✔ **Answer:** 0.793 **or** 0.821

0.821

**Explanation**
The accuracy can be found using the following code:

```
sex_class_model <- ifelse(test_set$Sex == "female" & test_set$Pclass != 3, 1, 0)
mean(sex_class_model == test_set$Survived)
```

Submit You have used 1 of 10 attempts

Answers are displayed within the problem

## Question 5a: Confusion matrix

2/2 points (graded)
Use the `confusionMatrix` function to create confusion matrices for the sex model, class model, and combined sex and class model. You will need to convert predictions and survival status to factors to use this function.

What is the "positive" class used to calculate confusion matrix metrics?

- [x] 0
- [ ] 1

✔

Which model has the highest sensitivity?

- [ ] sex only
- [ ] class only
- [x] sex and class combined

✔

Which model has the highest specificity?

- [x] sex only
- [ ] class only
- [ ] sex and class combined

✔

Which model has the highest balanced accuracy?

- [x] sex only
- [ ] class only
- [ ] sex and class combined

✔

Submit

You have used 1 of 2 attempts

Answers are displayed within the problem

## Question 5b: Confusion matrix

1/1 point (graded)
What is the maximum value of balanced accuracy?

0.806 ✔ **Answer:** 0.806 **or** 0.791

0.806

**Explanation**
The confusion matrix for each model can be calculated using the following code:

```
confusionMatrix(data = factor(sex_model), reference = factor(test_set$Survived))
confusionMatrix(data = factor(class_model), reference = factor(test_set$Survived))
confusionMatrix(data = factor(sex_class_model), reference = factor(test_set$Survived))
```

Submit You have used 1 of 10 attempts

Answers are displayed within the problem

## Question 6: F1 scores

2/2 points (graded)
Use the `F_meas` function to calculate $F_1$ scores for the sex model, class model, and combined sex and class model. You will need to convert predictions to factors to use this function.

Which model has the highest $F_1$ score?

- sex only
- class only
- sex and class combined ✔

What is the maximum value of the $F_1$ score?

0.872

✔ **Answer:** 0.855 **or** 0.872

$0.872$

**Explanation**

The $F_1$ score for each model can be calculated using the following code:

```
F_meas(data = factor(sex_model), reference = test_set$Survived)
F_meas(data = factor(class_model), reference = test_set$Survived)
F_meas(data = factor(sex_class_model), reference = test_set$Survived)
```

Submit

You have used 1 of 10 attempts

Answers are displayed within the problem

Ask your questions or make your comments about Titanic Exercises, part 1 here! **Remember, one of the best ways to reinforce your own learning is by explaining something to someone else, so we encourage you to answer each other's questions (without giving away the answers, of course).**

Some reminders:

- Search the discussion board before posting to see if someone else has asked the same thing before asking a new question.
- Please be specific in the title and body of your post regarding which question you're asking about to facilitate answering your question.
- Posting snippets of code is okay, but posting full code solutions is not.
- If you do post snippets of code, please format it as code for readability. If you're not sure how to do this, there are instructions in a pinned post in the "general" discussion forum.

## Discussion: Titanic Exercises, part 1

Show Discussion

**Topic:** Section 5: Classification with more than two classes and the caret package / 5.3.1: Titanic Exercises, part 1