uc3m | Universidad Carlos III de Madrid

Universidad Carlos III de Madrid

Introduction to Data Science

Second assignment: Machine Learning Models

Description

The objective of this assignment is to evaluate your data scientist skills in a realistic scenario.

Data

You will be working with the same data set used for the first assignment, which is included in the file **titanic_train.RDATA**. This file contains 668 observations on 10 columns:

- **Survived:** (1) or died (0).
- **Pclass:** Passenger's class (1st, 2nd, or 3rd)
- **Sex:** Passenger's age
- **Age:** Passenger's age
- **SibSp:** Number of siblings/spuoses aboard the Titanic
- **Parch:** Number of parents/children abooard the Titanic
- **Ticket:** Ticket number
- **Fare:** Fare paid for ticket
- **Cabin:** Cabin number
- **Embarked:** Location where passenger embarked on the ship (C-Cherbourg, S-Souththantom, Q-Queenstown)

For testing purposes, we kept a sufficiently large data set with the same format, representing future observations, however they are not included in the file (you do not have access).

Instructions

The objectives are twofold:

1. Using a supervised machine learning technique from those reviewed during the course, try to further analyze additional relationships among variables, which might influence survival (**survived**). This analysis complements the first assignment on exploratory data analysis, and it might help to support some of your previous conclusions, and also provide additional information which was difficult to extract manually.
2. You have to make a model for predicting variable "**survived**" as a function of the rest of variables included in the data set, or a subset.

Source code

Apart from all the data analysis that you think it is required to fulfill your task, your source code should define a function "**my_model**" that receives as input a *data.frame* with the same variables as the original dataset and returns a vector of predictions, with the same length as the input data. For example:

```
my_model = function(test){
  # do something
  pred = predict(my.fitted.model, test, type = "class")
  return(pred)
}
```

This function will make us easier to evaluate your model using our secret test data set.

Written report

You must hand out a report addressing the following points:

- Explain what technique from those reviewed in class you used to extract hidden relationships useful to predict survival (**survived**) based on the rest of variables, and the results of your analysis.
- The reason why you selected the proposed model and the parameter values. You should compare among different classifiers reviewed in class and explain your final selection.
- Do not forget to include an appropriate estimation of the performance of your model.

Additionally,

- You can make any transformation that you think is required on the data, as long as you give a feasible explanation for it.
- You can use **ggplot2** to support your analysis or any R package.

Evaluation

You must upload the following files:

- Source code
- Written report in PDF format (0 ~ 10 pages)
- Same groups as those in the first assignment are required. In case two individual groups during first assignment want to join for this one, please contact prof. Mínguez by email and let him know.
- There is going to be an oral presentation of your work with the following characteristics:
    - Maximum 4 slides (2 slides for each member of the group)
    - 10 minutes total (5 minutes each participant)

Due date November 24th at 22:00