Titanic - Machine Learning from Disaster

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

Titanic - Machine Learning from Disaster

Submit Prediction

Dataset Description

Overview

The data has been split into two groups:

- training set (train.csv)
- test set (test.csv)

The training set should be used to build your machine learning models. For the training set, we provide the outcome (also known as the “ground truth”) for each passenger. Your model will be based on “features” like passengers’ gender and class. You can also use feature engineering to create new features.

The test set should be used to see how well your model performs on unseen data. For the test set, we do not provide the ground truth for each passenger. It is your job to predict these outcomes. For each passenger in the test set, use the model you trained to predict whether or not they survived the sinking of the Titanic.

We also include **gender_submission.csv**, a set of predictions that assume all and only female passengers survive, as an example of what a submission file should look like.

Data Dictionary

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	

sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

Variable Notes

pclass: A proxy for socio-economic status (SES)

1st = Upper

2nd = Middle

3rd = Lower

age: Age is fractional if less than 1. If the age is estimated, is it in the form of xx.5

sibsp: The dataset defines family relations in this way...
Sibling = brother, sister, stepbrother, stepsister
Spouse = husband, wife (mistresses and fiancés were ignored)

parch: The dataset defines family relations in this way...
Parent = mother, father
Child = daughter, son, stepdaughter, stepson
Some children travelled only with a nanny, therefore parch=0 for them.

Files

3 files

Size

93.08 kB

Type

csv

License

Subject to Competition Rules

gender_submission.csv(3.26 kB)

2 of 2 columns

About this file

An example of what a submission file should look like.

These predictions assume only female passengers survive.

PassengerId

Survived

Label	Count
892.00 - 933.70	42
933.70 - 975.40	42
975.40 - 1017.10	42
1017.10 - 1058.80	41
1058.80 - 1100.50	42

1100.50 - 1142.20 42

1142.20 - 1183.90 41

1183.90 - 1225.60 42

1225.60 - 1267.30 42

1267.30 - 1309.00 42

892

1309

Label	Count
0.00 - 0.10	266
0.90 - 1.00	152

0

1

892089318940895089618970898189909001901090209030904190509061907190809090910191119120913091419150916191709181919092009210922092309241925192609270928192919300931093209330934093519361937093809390940194 11

Data Explorer

93.08 kB

gender_submission.csv

test.csv

train.csv

Summary

3 files

25 columns

Download All

Download data

Metadata

License

Subject to Competition Rules