Getting Started Prediction Competition

Titanic: Machine Learning from Disaster

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

Kaggle · 15,860 teams · Ongoing

Data 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.

Data (34 KB)

API `kaggle competitions download -c titanic` ? Download All

Data Sources

- gender_submission... 2 columns
- test.csv 11 columns
- train.csv 12 columns

About this file

An example of what a submission file should look like.

These predictions assume only female passengers survive.

Columns

- PassengerId
- # Survived

gender_submission.csv (3.18 KB)

2 of 2 columns Views

	PassengerId	Survived
	892 – 1309	0 – 1
1	892	0
2	893	1
3	894	0
4	895	0
5	896	1
6	897	0
7	898	1
8	899	0
9	900	1
10	901	0
11	902	0
12	903	0

	PassengerId	Survived
	 892 1309	 0 1
13	904	1
14	905	0
15	906	1
16	907	1
17	908	0
18	909	0
19	910	1
20	911	1

/