Overview

**Description**

Start here if...

You're new to data science and machine learning, or looking for a simple intro to the Kaggle prediction competitions.

Competition Description

The sinking of the RMS Titanic is one of the most infamous shipwrecks in history.  On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. This sensational tragedy shocked the international community and led to better safety regulations for ships.

One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others, such as women, children, and the upper-class.

In this challenge, we ask you to complete the analysis of what sorts of people were likely to survive. In particular, we ask you to apply the tools of machine learning to predict which passengers survived the tragedy.

Practice Skills

* Binary classification
* Python and R basics

**Evaluation**

Goal

It is your job to predict if a passenger survived the sinking of the Titanic or not.   
For each PassengerId in the test set, you must predict a 0 or 1 value for the Survived variable.

Metric

Your score is the percentage of passengers you correctly predict. This is known simply as ["accuracy”](https://en.wikipedia.org/wiki/Accuracy_and_precision#In_binary_classification).

Submission File Format

You should submit a csv file with exactly 418 entries plus a header row. Your submission will show an error if you have extra columns (beyond PassengerId and Survived) or rows.  
  
The file should have exactly 2 columns:

* PassengerId (sorted in any order)
* Survived (contains your binary predictions: 1 for survived, 0 for deceased)

PassengerId,Survived  
 892,0  
 893,1  
 894,0  
 Etc.

You can download an example submission file (gender\_submission.csv) on the [Data page](https://www.kaggle.com/c/titanic/data).

**Frequently Asked Questions**

### What is a Getting Started competition?

Getting Started competitions were created by Kaggle data scientists for people who have little to no machine learning background. They are a great place to begin if you are new to data science or just finished a MOOC and want to get involved in Kaggle.

Getting Started competitions are a non-competitive way to get familiar with Kaggle’s platform, learn basic machine learning concepts, and start meeting people in the community. They have no cash prize and are on a rolling timeline.

### What’s the difference between a private and public leaderboard?

The Kaggle leaderboard has a public and private component to prevent participants from “overfitting” to the leaderboard. If your model is “overfit” to a dataset then it is not generalizable outside of the dataset you trained it on. This means that your model would have low accuracy on another sample of data taken from a similar dataset.

Public Leaderboard

For all participants, the same 50% of predictions from the test set are assigned to the public leaderboard. The score you see on the public leaderboard reflects your model’s accuracy on this portion of the test set.

Private Leaderboard

The other 50% of predictions from the test set are assigned to the private leaderboard. The private leaderboard is not visible to participants until the competition has concluded. At the end of a competition, we will reveal the private leaderboard so you can see your score on the other 50% of the test data. The scores on the private leaderboard are used to determine the competition winners. Getting Started competitions are run on a rolling timeline so the private leaderboard is never revealed.

### How do I create and manage a team?

When you accept the competition rules, a team will be created for you. You can invite others to your team, accept a merger with another team, and update basic information like team name by going to the More < [Team](https://www.kaggle.com/c/titanic/team) page.

We've heard from many Kagglers that teaming up is the best way to learn new skills AND have fun. If you don't have a teammate already, consider asking if anyone wants to team up in the [discussion forum](https://www.kaggle.com/c/titanic/discussion).

### What are kernels?

Kaggle Kernels is a cloud computational environment that enables reproducible and collaborative analysis. Kernels supports scripts in R and Python, Jupyter Notebooks, and RMarkdown reports. Go to the [Kernels tab](https://www.kaggle.com/c/titanic/kernels) to view all of the publicly shared code on this competition. For more on how to use Kernels to learn data science, visit the [Tutorials tab](https://www.kaggle.com/c/titanic#tutorials).

### Why did my team disappear from the leaderboard?

To keep with the spirit of getting-started competitions, we have implemented a two month rolling window on submissions. Once a submission is more than two months old, it will be invalidated and no longer count towards the leaderboard.

If your team has no submissions in the previous two months, the team will also drop from the leaderboard. This will keep the leaderboard at a manageable size, freshen it up, and prevent newcomers from getting lost in a sea of abandoned scores.

*"I worked so hard to get that score! Give it back!"* Read more about our decision to implement a rolling leaderboard [here](https://www.kaggle.com/c/titanic/discussion/6240).

### How do I contact Support?

Kaggle does not have a dedicated support team so you’ll typically find that you receive a response more quickly by asking your question in the appropriate forum. (For this competition, you’ll want to use the [Titanic discussion forum)](https://www.kaggle.com/c/titanic/discussion).

Support is only able to help with issues that are being experienced by all participants. Before contacting support, please check the discussion forum for information on your problem. If you can’t find it, you can post your problem in the forum so a fellow participant or a Kaggle team member can provide help. The forums are full of useful information on the data, metric, and different approaches. We encourage you to use the forums often. If you share your knowledge, you'll find that others will share a lot in turn!

If your problem persists or it seems to be effective all participants then please [contact us](https://www.kaggle.com/contact).

**Tutorials**

Learning with Kaggle Kernels

Kaggle Kernels is an in-browser computational environment that is fully integrated with most competition datasets. Kernels is preloaded with most data science packages and libraries. It supports scripts and Jupyter Notebooks in R and Python, as well as RMarkdown reports. You can create submission files with Kernels and also use it to explore the competition data.

To get started with Kernels you can either:

1. Create a new script or notebook on the [Kernels](https://www.kaggle.com/c/titanic/kernels) tab or
2. “Fork” any kernel to create an editable copy for you to experiment with

We've selected some of the best kernels to help you get started with the competition. You can use the below kernels to create a submission file or to explore the data. Simply open the script or notebook and click "fork" to create an editable copy.

Getting Started with Python

Start with this easy-to-follow approach to using popular Python modules:

[Titanic Data Science Solutions Python Notebook](https://www.kaggle.com/startupsci/titanic/titanic-data-science-solutions)

* Use [pandas](http://pandas.pydata.org/) for data manipulation
* Use [matplotlib](http://matplotlib.org/) and [seaborn](http://seaborn.pydata.org/) for data visualization
* Learn to build models with [scikit-learn](http://blog.kaggle.com/author/kevin-markham/)

These Python kernels cover more advanced techniques and complex approaches:

[An Interactive Data Science Tutorial](https://www.kaggle.com/helgejo/titanic/an-interactive-data-science-tutorial)

* Get familiar with using Jupyter notebooks
* Learn the importance of feature selection in machine learning

[Machine Learning from Start to Finish with Scikit-Learn](https://www.kaggle.com/jeffd23/titanic/scikit-learn-ml-from-start-to-finish)

* Use cross-validation to make sure your model generalizes to new data (i.e., it doesn’t “overfit”)
* Use parameter tuning and grid search to select the best performing model out of several different classification algorithms

[XGBoost Example](https://www.kaggle.com/datacanary/titanic/xgboost-example-python)

* Learn how to the extremely popular [XGBoost](https://github.com/dmlc/xgboost) algorithm
* Click “show more” on the code tab to study the script

[An Introduction to Ensembling/Stacking in Python](https://www.kaggle.com/arthurtok/titanic/introduction-to-ensembling-stacking-in-python):

* Use the fundamental skill of “[ensembling](http://mlwave.com/kaggle-ensembling-guide/)” to combine the predictions of several models

Getting Started with R

[Exploring Survival on the Titanic](https://www.kaggle.com/mrisdal/titanic/exploring-survival-on-the-titanic)

* The basics of feature engineering and data visualization
* How to deal with missing values in the dataset
* How to train a random forest classifier to make a prediction
* If you’re not familiar with Rmarkdown, click on the “Code” tab to see the underlying code

[Families are Not Good For Survival](https://www.kaggle.com/jasonm/titanic/large-families-not-good-for-survival)

* Learn techniques to understand how your models are making predictions
* Use a visualization of a decision tree algorithm to compare different models
* Determine how features contribute to prediction accuracy

External Tutorials

* R & Python (interactive): Free, interactive tutorials that walk you through creating your first Titanic competition submission file are available from DataCamp ([R](https://www.datacamp.com/community/open-courses/kaggle-tutorial-on-machine-learing-the-sinking-of-the-titanic#gs.OAowLPw) / [Python](https://www.datacamp.com/community/open-courses/kaggle-python-tutorial-on-machine-learning#gs.=IXCXgA)) and Dataquest.io ([Python](https://www.dataquest.io/mission/74/getting-started-with-kaggle)). These tutorials are intended for those new to both machine learning and R or Python.
* R (local): A Kaggler created [tutorial](http://trevorstephens.com/kaggle-titanic-tutorial/getting-started-with-r/) that walks you through how to install R on your local machine and create a first submission.
* Excel: A [tutorial](https://www.kaggle.com/c/titanic/discussion/28323) on basic machine learning concepts in a familiar tool.