Trevor Stephens
Regular Data Scientist, Occasional Blogger.

Follow

Titanic: Getting Started With R

3 minutes read

So you're excited to get into prediction and like the look of Kaggle's excellent getting started competition, Titanic: Machine Learning from Disaster? Great! It's a wonderful entry-point to machine learning with a manageably small but very interesting dataset with easily understood variables.

In this competition, you must predict the fate of the passengers aboard the RMS Titanic, which famously sank in the Atlantic ocean during its maiden voyage from the UK to New York City after colliding with an iceberg.

Queenstown
Southampton
Cherbourg
New York
Titanic

While there could hardly be a more chaotic event than frightened people scrambling to escape a sinking ship, the disaster is famous for saving "women and children first". With an inadequate number of lifeboats available only a fraction of the passengers survived, and through this series of lessons, we'll try to predict who they were.

As with most Kaggle competitions, you are given two datasets:

- a training set, complete with the outcome (or target variable) for a group of passengers as well as a collection of other parameters such as their age, gender, etc. This is the dataset on which you must train your predictive model.
- a test set, for which you must predict the now unknown target variable based on the other passenger attributes that are provided for both datasets.

As this is a beginner's competition, Kaggle has provided a couple of excellent tutorials to get you moving in the right direction, one in Excel, and another using more powerful tools in the Python programming language. Ah, but you would feel (justifiably) embarrassed to use Excel, and Python seems a little heavy right now? Well you've come to the right place. I'm going to introduce you to a free and powerful statistical programming language called R and get you started with predictive analytics.

Over the next few weeks I'll ease you into R and its syntax, piece-by-piece, and step you through a selection of algorithms, from the trivial to the powerful. I'll also introduce some feature engineering concepts that will start to push the envelope.

In fact, by the time we're done, you'll have achieved big gains over the rest of the leaderboard by increasing your accuracy by only a few extra percentage points. That alone is a good lesson for Kaggle: those few points, or even fractions thereof, can translate to massive ranking swings and mean the difference between getting a top 10% badge on your profile (or even getting paid), once you're ready for the big leagues.

The guide is intended for people with zero experience in R, and probably very little programming experience as well. I won't get to cover all the syntax, but if you get through the lessons, you may wish to expand your horizons further by checking out some more broad tutorials here and here. Or if you're more of a book person, this is one that I can recommend highly: The Art of R Programming: A Tour of Statistical Software Design.

If you have any questions about these lessons, I'd encourage you to post them to the Kaggle forums where many others may have already come across the issue before and can jump in to help you out. If you notice any bugs or typos, or have any suggestions on making the tutorial easier to follow, please send me a direct message through Twitter. All code is available on my Github repository.

I will be dividing this series of tutorials into five parts:

- Part 1: Booting Up R
- Part 2: The Gender-Class Model
- Part 3: Decision Trees
- Part 4: Feature Engineering
- Part 5: Random Forests

So go ahead and get started with part 1

Tags: Competitions Tutorial

Categories: Kaggle-Titanic-Tutorial

Posted: January 10, 2014

Previous Next

LEAVE A COMMENT

29 Comments Trevor Stephens 1 Login

♡ Recommend 29 Share Sort by Best

Join the discussion…

LOG IN WITH

OR SIGN UP WITH DISQUS ?

Name

Bruno Barros • 4 years ago

Very good explanation about R and the Titanic's problem. Thank you Trevor.

29 ^ | ˅ • Reply • Share ›

Gilad Barkan • 4 years ago

you should be blessed for this gentle and didactic walkthrough

23 ^ | v • Reply • Share ›

Trevor Stephens Mod ➦ Gilad Barkan • 4 years ago

Thank you!

5 ^ | v • Reply • Share ›

Allhell Looz • 3 years ago

Hi Trevor .... I just started learning R with your post ... its pretty good !!

Can you point out good tutorials for Regression , classification , LDA , Clustering with R studio Or R-base with graph plotting and evaluation methods like you did here ??

Or Can you suggest any book with practical examples ??

Or can you write blog for that yourself :D

22 ^ | v • Reply • Share ›

A K R • 3 years ago

Great tutorial! ..thank you master!!!

12 ^ | v • Reply • Share ›

Shayan • 4 years ago

Fantastic tutorial!, Helped me so much as a newbie ^_^

12 ^ | v • Reply • Share ›

Josh • 4 years ago

Great tutorial! I am new to R and kaggle and this is just what I was looking for...thanks! Looking forward to Random Forests.

7 ^ | v • Reply • Share ›

Trevor Stephens Mod ➦ Josh • 4 years ago

Glad it helped!

2 ^ | v • Reply • Share ›

dnaeye • 4 years ago

OMG...I love your presentation. You should be a goddamn instructor/writer - if you aren't already.

6 ^ | v • Reply • Share ›

Yifan Xie • 4 years ago

Trevor, just to say a giant thank you to you! I am a newbie to statistic, so this well-structured tutorial comes in really handy, as I venture into predictive modelling and machine learning this year ^_^

6 ^ | v • Reply • Share ›

Trevor Stephens Mod ➦ Yifan Xie • 4 years ago

Glad it helped you! Good luck with the year to come!

2 ^ | v • Reply • Share ›

Manisha Rathore • 9 months ago

Thanks a lot sir,really helped a lot.I had no resources to learn this.

3 ^ | v • Reply • Share ›

Aragonda Jyosna • 7 months ago

Hi Trevor Excellent tutorial for beginners in ML.

I am getting the below output for this part of code :

```
> summary(combi$Embarked)
Length Class Mode
1309 character character
```

```
> which(combi$Embarked == '')
integer(0)
```

```
> factor(which(combi$Embarked == ''))
factor(0)
Levels:
Warning messages:
1: Unknown or uninitialised column: 'FamilyID2'.
2: Unknown or uninitialised column: 'FamilyID2'.
> combi$Embarked[c(62,830)] = "S"
Warning messages:
1: Unknown or uninitialised column: 'FamilyID2'.
2: Unknown or uninitialised column: 'FamilyID2'.
3: Unknown or uninitialised column: 'FamilyID2'.
```

Please let me know what I have to do in order to resolve this.

As I am unable to proceed further.

2 ^ | ˅ • Reply • Share ›

NoiseSignal • 6 months ago

This is a very helpful article Trevor, I learned a lot. Many thanks!

1 ^ | ˅ • Reply • Share ›

Aragonda Jyosna • 7 months ago

Hi Trevor Excellent tutorial for beginners in ML.

1 ^ | ˅ • Reply • Share ›

Annabelle Dsouza • 8 months ago

I'm a beginner at this. I know nothing machine learning, but I have basic knowledge of Java and databases and stuff. Is this tutorial appropriate for me?

1 ^ | ˅ • Reply • Share ›

Phenyo Molefe • 8 months ago

I am new here but am absolutely excited about the prospect of working through the matter provided above. I will furnish you with a review once complete. Thank you very much Trevor.

1 ^ | ˅ • Reply • Share ›

sebasluke • 10 months ago

Who rules??

YOU RULE!!!

1 ^ | ˅ • Reply • Share ›

mttt • a year ago

Thank you so much for this wonderful tutorial! This is so helpful!! Thanks again.

1 ^ | ˅ • Reply • Share ›

Rizkyk • 2 years ago

Trevor, just to say a giant thank you to you! I am a newbie to statistic, machine learning and other data scientist stuff. Your tutorial so well-structured comes in really handy, as I venture into predictive modelling and machine learning through year

1 ^ | ˅ • Reply • Share ›

jianyu chen • 3 years ago

Amazing tutorial! This will be a great start of Kaggle competition and R. Thank you!

1 ^ | ˅ • Reply • Share ›

Sonia Singh • 3 years ago

Thanks Trevor! :)

1 ^ | ˅ • Reply • Share ›

neel mani • 3 years ago

Hey Trevor!! Thanks for such an interesting guide for the newbie like me.

1 ^ | ˅ • Reply • Share ›

Mehdi Mujtaba • 3 years ago

Astounding work.Can we please get more?

1 ⌃ | ⌄ • Reply • Share ›

sachin_mk • 3 years ago

Just the thing I was looking for. Thanks !

1 ⌃ | ⌄ • Reply • Share ›

Andrey • 4 years ago

Thanks for the tutorial!

1 ⌃ | ⌄ • Reply • Share ›

Asmita Chatterjee • a month ago

hello friends , can any one suggest , where do we get the training and testing data set excel files ? I am not able to locate it ..

⌃ | ⌄ • Reply • Share ›

Leo R • a month ago

Hello, where can I find the same tutorial in python? or is it just for R?

⌃ | ⌄ • Reply • Share ›

Muhammad Habibi • 3 months ago

I like how you do story-telling.. make newbie become calm for what they don't know but become more curious in the same time.. I believe you're a good teacher/ tutor.. thanks!

⌃ | ⌄ • Reply • Share ›

ALSO ON **TREVOR STEPHENS**

Box-Plots for Education Recap

2 comments • 3 years ago

Trevor Stephens — Nope, not even close :-) ... I believe the test/train split had different schools or source systems …

Titanic: Getting Started With R - Addendum & Chocolate

7 comments • 4 years ago

Triyansha Vijay — Thank you for all these insights, and trivia. Its fun to read all this.

Titanic: Getting Started With R - Part 5: Random Forests

240 comments • 4 years ago

Raghavendra Saralaya — NVM, I had to make a new script with only the relevant code for part 5 - stuff from part 4 was …

Committed to Open Source… Again

5 comments • 3 years ago

Trevor Stephens — Thanks! It's in v0.17 now Kevin FYI

✉ Subscribe   Add Disqus to your siteAdd DisqusAdd   🔒 Privacy   DISQUS