

教程 | Kaggle初学者五步入门指南，七大诀窍助你享受竞赛

2017-07-22 机器之心

选自EliteDataScience

机器之心编译

参与：Panda、黄小天

Kaggle 是一个流行的数据科学竞赛平台，已被谷歌收购，参阅《业界 | 谷歌云官方正式宣布收购数据科学社区 Kaggle》。作为一个竞赛平台，Kaggle 对于初学者来说可能有些难度。毕竟其中的一些竞赛有高达 100 万美元的奖金池和数百位参赛者。顶级的团队在处理机场安全提升或卫星数据分析等任务上拥有数十年积累的经验。为了帮助初学者入门 Kaggle，EliteDataScience 近日发表了一篇入门介绍文章，解答了一些初学者最常遇到的问题。机器之心对这篇文章进行了编译介绍，另外也增加了一些机器之心之前发过的文章作为补充资源。


12 active competitions

Sort by Prize

Active All Entered Unlaunched

All Categories

Search




Passenger Screening Algorithm Challenge

Improve the accuracy of the Department of Homeland Security's threat recognition algorithms

Featured · 5 months to go

\$1,500,000

89 teams




Zillow Prize: Zillow's Home Value Prediction (Zestimate)

Can you improve the algorithm that changed the world of real estate?

Featured · 6 months to go

\$1,200,000

1,425 teams




Planet: Understanding the Amazon from Space

Use satellite data to track the human footprint in the Amazon rainforest

Featured · 9 days to go

\$60,000

840 teams



Instacart Market Basket Analysis

Which products will an Instacart consumer purchase again?

Featured · a month to go

\$25,000

1,307 teams

一些初学者会犹豫要不要参加 Kaggle 竞赛，这并不让人奇怪，他们通常有以下顾虑：

- 我该如何开始？
- 我要和经验丰富的博士研究生比赛吗？
- 如果没有获胜的机会，还值得参与吗？
- 这就是数据科学吗？（如果我在 Kaggle 上表现不好，我在数据科学领域还有希望吗？）
- 未来我该如何提升我的排名？

如果你有其中任何问题，你就看对了文章。在这篇指南中，我们会解读上手 Kaggle、提升技能和享受 Kaggle 所需要了解的一切。



Kaggle vs. 「经典的」数据科学

首先，我们要清楚了解：

Kaggle 竞赛和「经典的」数据科学有一些重要的不同之处，但只要你以正确的心态接触它，就也能收获有价值的经验。

让我们解释一下：

Kaggle 竞赛

本质上，带有奖金池的竞赛必须满足一些标准：

- 问题必须困难：竞赛不应该是一个下午就能解决的任务。为了得到最好的投资回报，主办公司会提交他们最大最难的问题。
- 解决方案必须新：要赢得最新的竞赛，你通常需要进行扩展研究、定制算法、训练先进的模型等等。
- 表现必须能比较：竞赛必须要决出优胜者，所以你和其他对手的解决方案必须要被评分。

「经典的」数据科学

相对而言，日常所用的数据科学并不需要满足这些标准。

- 问题可能简单。实际上，数据科学家应该尽力确认易于实现的成果：可以快速解决的富有成效的项目。
- 解决方案可以是成熟的。大多数常见任务（比如探索分析、数据清理、A/B 测试、经典算法）都已经有了已得到证明的框架。没必要重新发明轮子。
- 表现可以是绝对的。即使一个解决方案只是简单地超越了之前的基准，那也非常有价值。

Kaggle 竞赛鼓励你竭尽所能，而经典数据科学则推崇效率和最大化的业务效果。

Kaggle 竞赛值得参加吗？

尽管 Kaggle 和经典数据科学之间存在差异，但 Kaggle 仍然是一种很好的入门工具。

每个竞赛都是独立的。无需设置项目范围然后收集数据，这让你有时间专注其它技能。

练习就是实践。学习数据科学的最好方法是在做中学。只要没有每场竞赛都获胜的压力，你就可以练习各种有趣的问题。

讨论和获胜者采访很有启发性。每个竞赛都有自己的讨论板块与获胜者简报。你可以窥见更有经验的数据科学家的思考过程。

The Nature Conservancy Fisheries Monitoring Competition, 1st Place Winner's Interview: Team 'Towards Robust-Optimal Learning of Learning'

Kaggle Team | 07.07.2017

This year, The Nature Conservancy Fisheries Monitoring competition challenged the Kaggle community to develop algorithms that automatically detects and classifies species of sea life that fishing boats catch. Illegal and unreported fishing practices threaten marine ecosystems. These algorithms would help increase The Nature Conservancy's capacity to analyze data from camera-based monitoring systems. In this winners' interview, first place team, 'Towards Robust-Optimal Learning of Learning' (Gediminas Pekšys, Ignas Namajūnas, Jonas Bialopetravičius), shares details of their approach like how they needed to have a ...

2017 Data Science Bowl, Predicting Lung Cancer: 2nd Place Solution Write-up, Daniel Hammack and Julian de Wit

Kaggle Team | 06.29.2017



Stacking Made Easy: An Introduction to StackNet by Competitions Grandmaster Marios Michailidis (KazAnova)

Kaggle 获胜者采访

怎样入门 Kaggle?

接下来，我们将给出一个按步进行的行动规划，然后慢慢上升到 Kaggle 竞赛中。

第一步：选择一种编程语言

首先，我们推荐你选择一种编程语言，并坚持使用。Python 和 R 在 Kaggle 和更广泛的数据科学社区上都很流行。

如果你是一个毫无经验的新手，我们推荐 Python，因为这是一种通用编程语言，你可以在整个流程中都使用它。

参考：

- 数据科学领域 R vs Python: <http://elitedatascience.com/r-vs-python-for-data-science>
- 如何为数据科学学习 Python: <http://elitedatascience.com/learn-python-for-data-science>
- 深度 | R vs Python: R 是现在最好的数据科学语言吗?

- [业界 | 超越 R，Python 成为最受欢迎的机器学习语言](#)

第二步：学习探索数据的基础

加载、浏览和绘制你的数据（即探索性分析）的能力是数据科学的第一步，因为它可以为你将在模型训练过程中做的各种决策提供信息。

如果你选择了 Python 路线，那么我们推荐你使用专门为这个目的设计的 Seaborn 库。其中有高层面的绘图函数，可以绘制许多最常见和有用的图表。

参考：

- Seaborn 库：<https://seaborn.pydata.org/>
- Python Seaborn 教程：<http://elitedatascience.com/python-seaborn-tutorial>
- [资源 | 2017 年最流行的 15 个数据科学 Python 库](#)

第三步：训练你的第一个机器学习模型

在进入 Kaggle 之前，我们推荐你先在更简单更容易管理的数据集上训练一个模型。这能让你熟悉机器学习库，为以后的工作做铺垫。

关键在于培养良好的习惯，比如将你的数据集分成独立的训练集和测试集，交叉验证避免过拟合以及使用合适的表现评价指标。

对于 Python，最好的通用机器学习库是 Scikit-Learn。

参考：

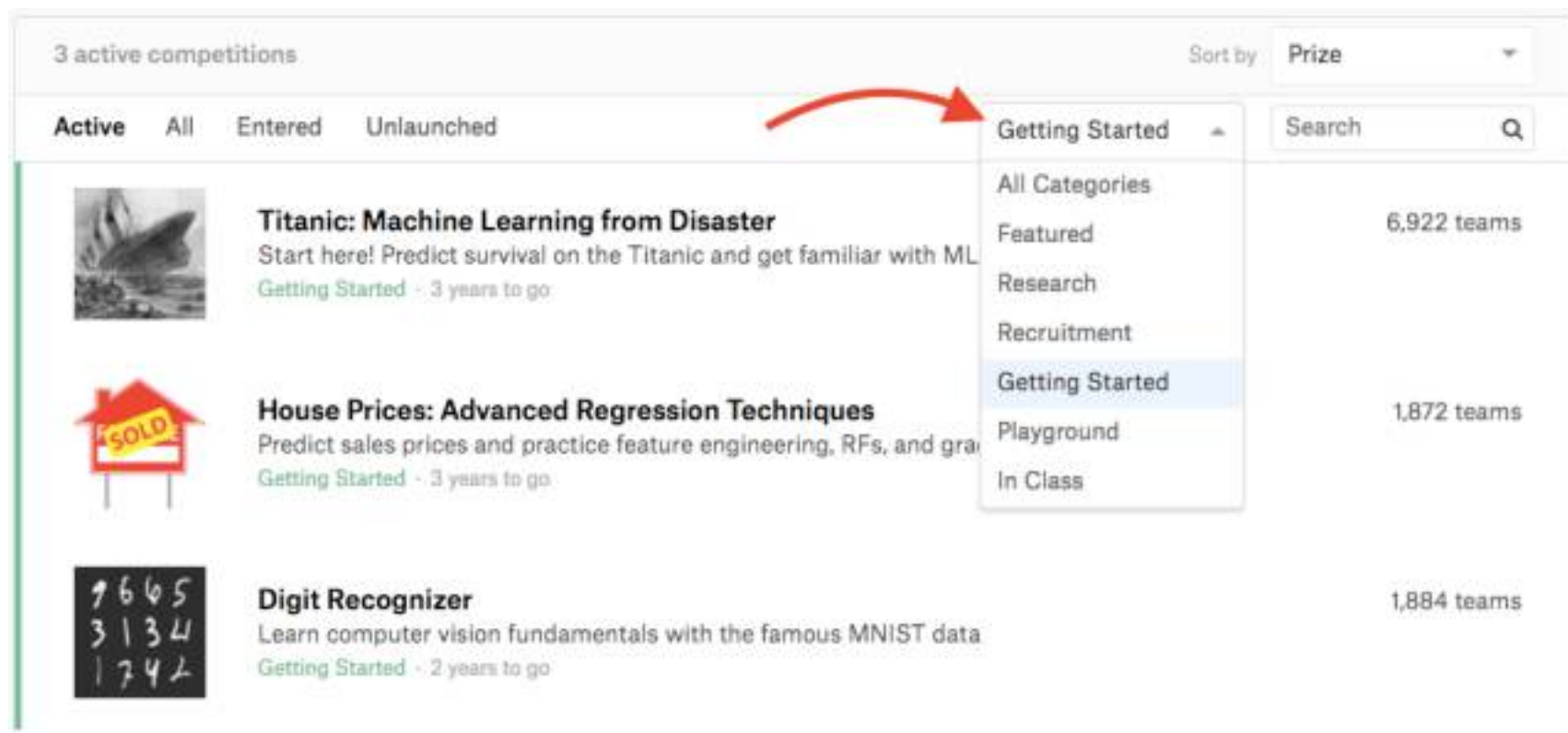
- Scikit-Learn 库：<http://scikit-learn.org/stable/>
- Python Scikit-Learn 教程：<http://elitedatascience.com/python-machine-learning-tutorial-scikit-learn>
- 7 天应用机器学习速成课：<http://elitedatascience.com/>
- [只需十四步：从零开始掌握 Python 机器学习（附资源）](#)

第四步：解决入门级竞赛

现在我们已经准备好尝试 Kaggle 竞赛了，这些竞赛分成几个类别。最常见的类别是：

- Featured：这些通常是由公司、组织甚至政府赞助的，奖金池最大。
- Research：这些是研究方向的竞赛，只有很少或没有奖金。它们也有非传统的提交流程。
- Recruitment：这些是由想要招聘数据科学家的公司赞助的。目前仍然相对少见。
- Getting Started：这些竞赛的结构和 Featured 竞赛类似，但没有奖金。它们有更简单的数据集、大量教程和滚动的提交窗口让你可以随时输入。

Getting Started 竞赛非常适合初学者，因为它们给你提供了低风险的学习环境，并且还有很多社区创造的教程：<https://www.kaggle.com/c/titanic#tutorials>



第五步：比赛是为了更好地学习，而不是赚钱

有了上面的基础，就可以参与到 Featured 竞赛中了。一般来说，为了取得好排名，通常需要远远更多的时间和精力。

因此，我们建议你明智地选择参与项目。参加竞赛能帮你深入到你希望长期参与的技术领域中。

尽管奖金很诱人，但更有价值（也更可靠）的回报是为你的未来事业所获得的技能。

享受 Kaggle 的小诀窍

最后，我们将介绍几个参与 Kaggle 的最受欢迎的诀窍，希望能帮你享受你的 Kaggle 时光。

诀窍 1：设置循序渐进的目标

如果你曾经玩过什么让人上瘾的游戏，你就知道循序渐进的目标的重要性。那就是好游戏让人着迷的诀窍。每一个目标都要足够大，以便带来成就感；但也不能太大，不然无法实现。



大多数 Kaggle 参与者都没赢过任何一场竞赛，这完全正常。如果把获胜作为第一个里程碑，你可能会失望，尝试几次之后可能就会失去动力。循序渐进的目标会让你的旅程更加愉快。比如：

提交一个超越基准解决方案的方案

- 在一场竞赛中进入排名前 50%
- 在一场竞赛中进入排名前 25%
- 在三场竞赛中进入排名前 25%

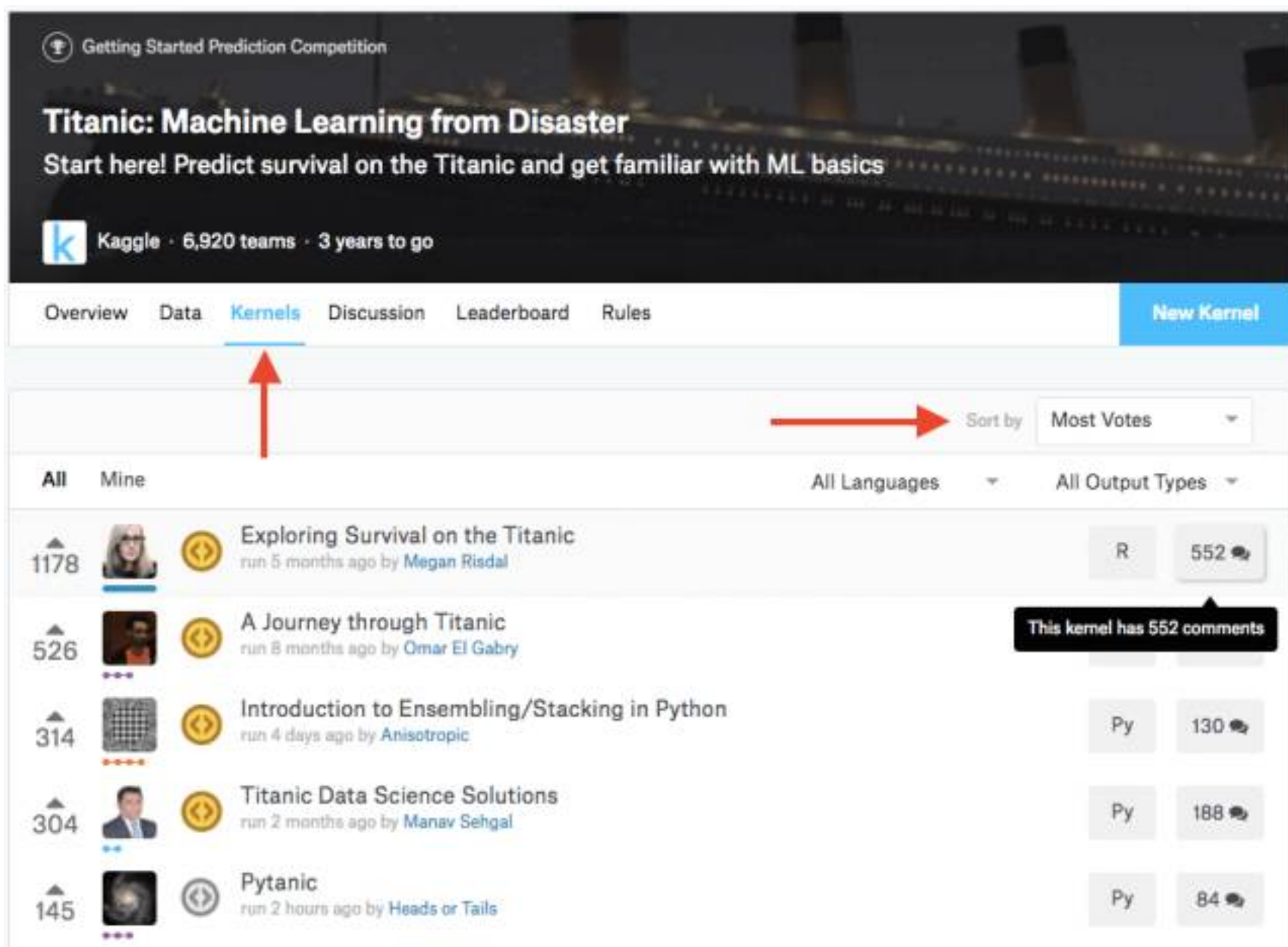
- 在一场竞赛中进入排名前 10%
- 赢得一场竞赛！

这种策略让你可以一路衡量你的进展和进步。

诀窍 2：查阅得票最多的 kernel

Kaggle 有一个非常厉害的功能：参与者可以提交 kernel，即用于探索一个概念、展示一种技术或分享一种解决方案的短脚本。

当你开始一场竞赛或感觉进步停滞时，查阅受欢迎的 kernel 或许能给你带来灵感。



诀窍 3：在论坛中提问

不要害怕问「愚蠢的」问题。

提问能遇到的最糟糕的事情是什么？也许你会被忽视.....仅此而已。

另一方面，你能得到很多回报，包括来自经验更丰富的数据科学家的建议和指导。

诀窍 4：独立发展核心技能

开始的时候，我们建议你独自工作。这将迫使你解决应用性机器学习流程中的每一步，包括探索性分析、数据清理、特征工程和模型训练。

如果过早地和人组队，你就可能会错失发展这些基本技能的机会。

诀窍 5：组队以拓展你的极限

虽然太早组队不好，但在未来的比赛中组队让你能向其他人学习，进而拓展你的极限。过去的许多获胜者都是团队，这让他们可以结合彼此的知识共同施展力量。

此外，一旦你掌握了机器学习的技术技能，你就可以与其他可能比你有更多领域知识的人合作，进一步扩展你的机遇。

诀窍 6：记住 Kaggle 可以成为你的垫脚石

记住，你不一定要成为一个长期的 Kaggle 人。如果发现你不喜欢这种形式，也没什么大不了的。

实际上，许多人在做自己的项目或成为全职数据科学家之前都会使用 Kaggle 作为自己的垫脚石。

所以你的关注重点应该是尽可能地学习。长远来看，参与能给你带来相关经验的竞赛比参加有最高奖金的竞赛更好。

诀窍 7：不要担心排名低

有些初学者担心低排名出现在他们的个人资料中，结果一直没有开始。当然，比赛焦虑是很正常的现象，并不只限于 Kaggle。

但是，排名低真的没什么关系。没人会因此贬低你，因为他们曾经某个时候也是初学者。



即便如此，如果仍然担心个人资料里的低排名，你可以再单独创建一个练习账号。一旦觉得自己能力不错了，就可以开始用你的「主帐号」来建立丰功伟绩了。（再说一下，这么做毫无必要！）

结论

在这篇指南中，我们分享了上手 Kaggle 的 5 大步骤：

1. 选择一种编程语言
2. 学习探索数据的基础
3. 训练第一个机器学习模型
4. 解决入门级竞赛
5. 比赛是为了更好地学习，而不是赚钱

最后，我们分享了享受这个平台的 7 个诀窍：

- 设置循序渐进的目标
- 查阅得票最多的 kernel
- 在论坛中提问
- 独立发展核心技能
- 组队以拓展你的极限

- 记住 Kaggle 可以成为你的垫脚石
- 不要担心排名低

原文链接：<https://elitedatascience.com/beginner-kaggle>

本文为机器之心编译，转载请联系本公众号获得授权。



加入机器之心（全职记者/实习生）：hr@jiqizhixin.com

投稿或寻求报道：editor@jiqizhixin.com

广告&商务合作：bd@jiqizhixin.com

点击阅读原文，查看机器之心官网↓↓↓

[阅读原文](#)