Kaggle首战金牌总结!

尤而小屋 2022-07-16 00:00 发表于北京

收录于合集
#kaggle 23 #机器学习 110 #比赛 2 #人工智能 38

今天给大家分享一篇大佬参加Kaggle的经验文章，作者是一名阿里算法工程师，希望对也想玩转Kaggle竞赛的朋友有所帮助，以下为原文。

> 作者：jiazhuamh
>
> 来自：https://zhuanlan.zhihu.com/p/60953933

这篇文章是我对自己第一次参加 kaggle 竞赛并获得金牌(14/4129)的一个总结，谈不上太多经验，涉及到的一些比赛规则和小技巧希望能对刚刚开始打 kaggle 比赛的小伙伴起到一些帮助。

一. 平台简介

kaggle 是全球首屈一指的数据科学、机器学习竞赛和分享平台。很多大公司作为出题方，会将问题和相关数据放在平台上形成一个竞赛，所有的 kaggle 用户都可以参加，获胜的团队或个人既能拿到奖金，又能获得奖牌对于新手还能收获实战经验。如果能在 kaggle 竞赛中获得一个不错的排名的话，对于自身履历或面试还是有很大帮助的。但由于竞争激烈，想在榜单上拿到一个较高的排名并不容易。

1.1 比赛介绍

kaggle 网站是纯英文的，刚开始的时候可能需要花点时间熟悉一下各板块，之后多数时间我们会待在 Competition 板块下，这个板块包含了所有与这次比赛相关的信息和操作。

1.2 Overview

介绍比赛的背景信息、结果的评估指标、比赛时间线和奖金。

刚参加一个比赛，需要花点时间了解这个比赛的领域背景，甚至需要查一些资料或阅读一些文献，这对后面构建特征和选择模型很重要。我看到有很多 winners 分享经验说自己构建的大多数特征都是从商业(领域)层面思考得到的，所以领域的先验知识很重要。

另一需要注意的是比赛的时间线。比赛有一个开始时间，一个组队截止时间和一个最终提交时间。一般一个比赛会持续几个月，最终提交时间就是比赛结束的标志。

组队截止时间一般是比赛结束前一周，过了这个时间点就不允许再组队了。留意好这些时间点，对你把握比赛的进度至关重要，尤其是用业余时间打比赛的上班族。很多比赛的数据量比较大，模型跑下来很耗时，如果到最后半个月才开始发力，会发现时间不够用，很多想法都没机会尝试。

有大牛分享自己的时间安排是：在比赛刚开始的时候会多花点时间做探索，把 pipeline 搭起来，接下来可以少花点时间，平时有什么想法可以直接测试，最后一个月或半个月再做集中冲刺。

1.3 Data

介绍数据，提供数据下载。

这个模块需要认值阅读，它介绍数据的产生方式、存储形式、每个字段的含义等。我们很多时候是通过对数据规模或形式的判断来决定要不要参加这个比赛。

比如，数据规模很大，你没有内存足够大的服务器可以hold住，可能就没法打这个比赛；再比如，是图像数据，那就得用深度神经网络，如果对这方面不熟或者没有GPU可用，可能也没法打这个比赛。对新手而言，该开始可能更倾向于择一些表格类、数据量适中的数据集。

1.4 Kernels

脚本区(不知道怎么翻译，大致是这个意思)。支持 Python 语言的脚本 .py 和 .ipynb，和 R 语言的脚本 .R 和 .ipynb。

分 public kernel 和 private kernel。public kernel是公开的，大家都可以看到，从这里可以学到非常多的东西，当然你自己也可以通过公开自己的 kernel 分享解决方案或观点。private kernel是你自己的，别人看不见，你可以分享给组内成员。

为方便大家打比赛，kaggle 提供了一些运算资源。kaggle 用户的每个 kernel 可以有 16G 的内存和 4 核CPU，这足够打多数比赛了。另外，最近还提供了 GPU，在新建 kernel 的时候可以选择开启 GPU，但当打开 GPU 时，CPU 和内存资源会少一些，大家可以根据情况选择使用。

国内用户可能会存在一个延迟或连接不稳定的问题，有时候甚至刷不出，这个应该是因为墙的原因，最好自己花钱搞个翻墙的代理。

1.5 Discussion

讨论区，这个区大家会分享观点、讨论问题、甚至寻找组队队友。

kaggle 的分享氛围非常好，在整个比赛进程中大家不断地分享自己的新发现，很多有用的信息都是在这里获取的。

对于一个新手而言，每天做好 kernel 区和 discussion区的跟踪，有充足的时间尝试他们的想法，应该可以获得一个不错的排名。比赛结束后，这个更是一个知识的盛宴，winners以及大牛会将自己获胜用到的方法、小技巧（tricks）全部分享出来。可以说，如果认真打完一场比赛，不论排名如何，对自己都是一个质的提升。

1.6 Leaderboard

排名区，分 public LB 和 private LB。比赛方会将 test 数据集中一部分(比如 30%)拿出来做为 public LB 评分和排名，剩下的部分作为 private LB（也就是最终结果）的评分和排名。

你每天都可以提交并查看自己的答案在 public LB 的得分和排名情况，在比赛结束前需要选择两个提交作为自己的最终答案，比赛结束后，平台会计算你的答案的 private LB 得分并自动挑选得分高的一个作为你的最终成绩。

在讨论区你会经常听到大家讨论 CV score、LB score，指的就是你模型本地交叉验证的得分和提交后的 public LB 得分。

需要注意的是，public LB 得分可能会和 private LB 得分差别很大，比赛结果公布前你可能排名前十，比赛结果公布后发现自己跌到上千名了，这就是所谓的 shake up，很吓人的。这次的 Elo 比赛因为异

常值的原因也有很大的 shake up，大家很早就预料到了，好在我们组做了一定的准备，抗住了 shake up，从第 8 名跌倒 14 名，还算是比较幸运的。

1.7 Rules

比赛规则。这个很容易被新手忽略，因为都是讲一些条款。很多新手甚至是大牛都在这个上面吃过亏。讲两个需要注意的地方。

一是提交次数，这里会写明每天允许的最大提交次数，一般是 5次，也有比赛时 3次的。假如一个比赛持续时间是三个月，那么总提交次数差不多就是 5×90=450 次，即便组队，你们对的总提交次数也是这个值。假如你们对的三个成员，在比赛最后一个月打算组队，大家已经各自提交了200、100、50次，是没法组队的，必须10天不提交，等提交次数达到 350（5×70天）。

很多人为了躲过提交次数的限制或者“节省提交次数”，专门注册了小号，这是很危险的，这被称为multiple accounts，是会被 kaggle 的反作弊系统侦察出来的。在比赛结束后，会先公布初步排名，然后 kaggle 平台反作弊系统开始运行，大约两三天后，凡是被判为作弊的队伍直接从排名中移除，几个月的努力就打水漂了！

另一个是组外私自分享代码和结果，这也是明令禁止的。组队之后队员之间可以分享，或者通过公开的kernel或discussion区分享。同样，如果被检测出不同队伍或个人间有相似的结果，也会被移除最终榜单。

1.8 Team

这里管理你的队伍。可以向别人发起组队邀请，或者接受别人的邀请，当然，也可以时不时给自己队伍改一个骚气的名字。我们队伍最后集齐四人之后，改名为 skynet，是电影《终结者》里人类创造的一个人工智能防御系统，本来是想通过这个霸气的名字“终结”其他人的，后来发现我们还是太年轻太幼稚了，差点被别人终结。

二. 挑选比赛

对于新手而言，我觉可以从一下几个方面考虑。

2.1 比赛类型

借用 Eureka 大牛的观点，可将 kaggle 平台上的比赛分成挖掘、图像、语音和 NLP 四类。其中挖掘类主要面对的是结构化数据，也就是表格数据，包括了各式各样的预测问题(预测销量、点击率、推荐排序等)。主要的共性就是理解数据，理解问题，从数据中找到有用的信息用来预测，这类问题胜负更多的是在特征上。

对于图像问题，可能就较少涉及到特征了。图像问题现在主要用到深度学习的相关技术，基于深度学习做了很多改进或者演绎，已经完全不需要在特征层面上去做什么了。

对于新手而言，挖掘类问题可能更容易上手。

2.2 数据规模

新手尽量选择中等规模的数据，这样数据处理和模型训练的时间更短，方便尝试尽可能多的方法。新手缺乏经验，可以通过多尝试来弥补，但如果选择的数据集太大，尝试的时间成本和计算成本都很大，不管从时间还是精力层面考虑，都是很大的障碍。

数据尽可能有丰富多样的特征，太完美的数据集不要选。这个大家可能会比较费解，完美的数据集不是更方便吗？没有缺失值，都是数值特征，多方便。

其实不然，越是这样的数据集，可供新手发挥的余地就越小，你会发现你唯一能做的就是把数据喂给模型，结果是怎样就是怎样了，甚至不知道如何改进、提升。所以，尽量选择有缺失值、有很多类别特征的数据集，这样你可以借鉴前人经验对这些特征做各种变换、尝试，时不时能收获一些小惊喜，甚至最后还能取得一个不错的成绩。

三. 如何打好一个比赛

起这样一个标题有点大言不惭，可能需要一个身经百战的 Grand Master 来回答这个问题。我这里想分享的是作为一个菜鸟，怎么一步步通过学习和尝试拿到自己的第一个奖牌。我就以这次参加的Elo Merchant Category Recommendation 为例讲讲自己的经历和感悟。

3.1 比赛早期加入

我是在比赛开始后不久就加入的，对于新手而言，早点加入比赛有好处，因为刚开始人不多，大牛更少(他们都是最后一个月强势加入然后霸占榜首的)，这样你每天的尝试都是有排名上的收获，这对自己是一个很大的鼓舞。

3.2 理解背景、探索数据与验证

比赛的背景我看了很多遍：Elo是一家信用公司，它提供了信用卡用户的一些信息和历史交易记录，想通过这些数据预测用户忠诚度。根据它的描述，我通过数据逐条验证，发现有疑问的地方就会在discussion区里提，也会发布一些公开的 kernel 验证自己的猜测。其中有两个比较重要的是 ① target 的预测偏差90%是异常值贡献的，② hist_trans 和 new_trans 的一些基本信息的验证。这些都很简单，但对后面的模型起到了很大的帮助。

3.3 零模型和基模型

这算是比赛的套路了，先训练一个baseline model, 然后在这个基础上不断优化、提升。我这里要补充的一个点是“零模型”，就是不用模型去做预测，一般就用全零或者全均值去做提交。比如这次的Elo比赛，用全零提交，public LB 得分3.9x，而基模型也只有 3.8x。通过这个对比可以看到模型的效果是很差的，这其实是太多异常值导致的。这些信息对后续的处理或模型评估都很有用。

GBDT已成为挖掘类比赛的主流算法，现成的包有xgboost、lightgbm、catboost，lightgbm因为速度快、准确度高脱颖而出。所以这次比赛，lightgbm 仍是主流的工具。

当然，前期也需要尽可能多地尝试其他模型，机器学习领域有一个“没有免费午餐定律”，是说没有那种模型要比其他模型都好，不同的数据集可能适合不同的模型，这个只用通过尝试才知道。所以，打kaggle比赛，其实更像是不断地做数据实验。我早期也试了**libffm、RandomForest、LinearRegression、DeepNN等。**

3.4 本地评估体系

模型评估体系没建起来，其他任何探索都是盲目的，因为你没法确定好坏。这个很有讲究，你得判断public LB 和 private LB 的得分是不是一致的或高度相关的。否则你可能在接下来的几个月中“一直努力拟合public LB”，结果却偏离了 private LB，这就是为何大牛一直告诫我们“Trust your CV”的原因。

这次Elo比赛，我早起发现了可以通过手动矫正异常值来提高 public LB 得分，后面通过统计分析与验证，发现原来是 public Lb 中异常点比随机水平高，这是很危险的，后来就采用了保守的做法，比赛结

果公布后证实了这个想法。

3.5 特征重要性和特征工程

前期适当做一些调参，获取一组相对可以的参数就可以了，不用太花太多时间调参，因为它起得作用很有限。用这组参数和基模型来输出特征重要性。挑选重要的特征(比如top10)进行分析和构建新特征，然后测试，然后扩大范围(比如top20)重复以上过程。

这一步是拉开差距的关键，也是花最多精力的地方。在这个过程中需要不断地关注discussion和kernel区，多看别人时怎么提取特征的，多看别人的 EDA，多尝试，说不定就有启发你的地方。

对于Elo比赛，前期大家模型输出的top importance特征都是时间相关的，虽然难以解释，但方向有了，所以我们也构建了很多时间相关的特征，事实证明，这些特征起到了重要的作用。直到后期，有位Grand Master利用他对数据神乎奇迹的转换和惊人的洞察力揭开了忠诚度的商业含义，时间相关特征为何如此重要才得到了解释。

3.6 组队

对于新手而言，你的排名就是你组队的筹码。我前期发现异常值矫正后，排名飙升到top20，就有很多人联系我组队了，我在discussion区与北京一个小伙伴组队了，没过多久又邀请了一名在新加坡读书的小伙伴，之后我们一直维持在top10的水平，一度长时间位于top5，差点迷失自己，直到后期大牛进入把我们挤出了top10，我们开始紧张起来，又吸收了一名韩国的朋友，慢慢回到了top10。

我们都是第一次打比赛，不过配合得挺好的，建了微信群用英文交流，几乎无障碍。

3.7 模型融合

特征工程和模型融合是打 kaggle 比赛的两大利器，比赛后期基本就是尝试各种模型融合了。对于模型融合，大牛的观点：真正重要的是要有两三个强模型和很多相关性很小的弱模型，模型多样性很重要，弱一点也没关系。幸运的是我们每个人手里都有一个不错的模型，然后我又批量生成了一些弱模型，融合起来，取得了不错的效果。

四. 结束语

总结得很不系统，想到哪里就写到哪里了，里面提到的东西可能大家早就知道了，希望能对新手有一些启发，更过干货可以直接上 kaggle 平台 discussion 区搜索 solutions，有数之不尽的大牛分享，干货满满，Happy Kaggling！、

推荐阅读

MySQL必须掌握4种语言!
三大树模型实战乳腺癌预测分类
Plotly+Pandas+Sklearn：实现用户聚类分群!
用户群组分析，Python实现!
Kaggle可视化：黑色星期五画像分析
kaggle实战：可视化深度探索苹果AppStores
kaggle实战：6大回归模型预测航班票价

> 尤而小屋，一个温馨的小屋。小屋主人，一手代码谋求生存，一手掌勺享受生活，欢迎你的光临

尤而小屋

尤而小屋，一个温馨且有爱的小屋🏠 小屋主人，一手代码谋求生存，一手掌勺享受…

261篇原创内容

公众号

收录于合集 #kaggle 23

上一篇
德国信贷数据建模baseline

下一篇
kaggle二分类问题：中风病人预测

喜欢此内容的人还喜欢

基于PySpark的机器学习入门!
尤而小屋