

一行代码逆转山东赛Top1

YueTan kaggle竞赛宝典 2022-03-31 23:30

↑↑↑关注后"星标"kaggle竞赛宝典


kaggle竞赛宝典技巧

作者: YueTan

一行代码逆转山东赛Top1

简介

上百数据实战技巧，欢迎关注Kaggle竞赛宝典。



kaggle竞赛宝典

数据竞赛Top方案，竞赛黑科技，竞赛到入职的一些感想。
272篇原创内容

公众号

近期看到YueTan的一篇文章，挺有意思，和大家分享一下。

现实真的不是拍电影

本次比赛，我实践的是中国古拳法的最高境界，心理战术。



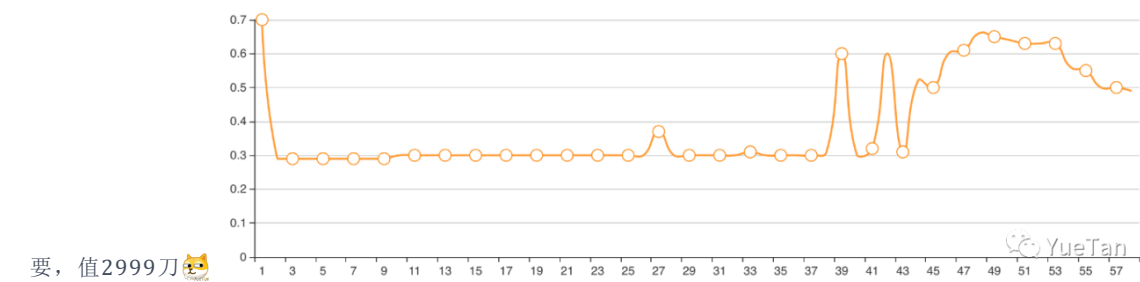
这次运气好，才能戏弄作弊组织罢了。不过，我也被组织戏弄了，大家扯平啦。我一直在说第3456有问题，当天我就被他们挤到了第3，我成替身了。但是，



真实情况是，比赛进行一周，我就借大佬的特征工程文章遥遥领先，尽在掌控之中。最后，有惊无险保住了位置。逆转的关键，一行代码：

```
# 藏分操作 :)
submit['ROOM_EMPTY'] = 1 - submit['ROOM_EMPTY']
```

把auc藏分的一行注释掉，分数立马涨到了第一。就像一个故事说的，这一行代码本身不重要，只值1刀；这一行代码写在我这里才重



要，值2999刀

全部代码已开源：github.com/LongxingTan/Data-competitions

就像总有人喜欢在比赛中作弊一样，我这次的藏分、运用失败的心理战术，都是游戏的一部分。也许本身并无对错高下之分，作弊与反作弊只是互有攻防。但是强中自有强中手，一山还有一山高，作弊手段过于低级，小心眼的我就会写成九集，每天不停连播。

真实招数

所以，这一行代码只是个假动作、标题党。真实操作还是一如既往的朴素，那就是“没有中间商赚差价”。由于民宿数据有较大质量问题，如何制定标签都不清晰。目标不确定的话，模型的学习效果就存疑了，所以我直接作了目标编码相关特征，当结果就提交了。

答辩尚未进行，参与答辩的选手请自觉关闭本文。别往下看了，互联网不是法外之地，回头是岸。

特征工程的艺术

与其说特征工程是一门工程技术，不如说她更像艺术。既有套路，更需要创造性。几十年前我上大学时，那时民风还很淳朴，一点不卷。所以我很闲暇地读过不下十遍梵高传记《渴望生活》，让我这个对艺术一窍不通的乡熊，略懂了一点何为一流的艺术。

乡熊 - 百度百科



乡熊，威海（山东最东端胶东半岛上）方言。“乡”意思是乡下的，没见过世面的，孤陋寡闻的。“熊”是方言发音的近似写法，与动物无大关系，意思是不本事，无能的，窝囊等。**乡熊**是威海人民经常用来自嘲的用语，因为上几代大部分还都是农民，这样说有种归属感。

百度百科

YueTan

艺术只是形式，表达自我才是本质，梵高的画里表达了他对世界的理解，这份理解，就像牛顿和麦克斯韦方程一样，都是对世界的创造性理解。不同的是，可以用方程表达，也可以用颜料表达。



回到机器学习，特征工程也是一种自我表达，表达的是对某个场景和业务的理解。（当然，买菜和借钱的场景不论理解的多深，相比科学或艺术的世界还是low了。但是给我一个low的机会，我肯定感激涕零啊）

回想初次参加比赛，我也是，把头发梳成大佬模样，照着大佬比赛开源，一股脑开始加特征，最后毫无成效。就像画手刚入行时，留长头发、穿奇怪衣服，试图让自己看起来像艺术家。慢慢才发现，这些不重要，真正的理解和表达才是关键。

所以在我看来，特征工程与科学、艺术一样，关键都在于**理解和表达**。回到大佬的文章，**理解是发现问题**，是从业务、先验、eda、实验结果中发现端倪，可以基于先验的经验积累，可以思考场景的独特性，可以借助模型特征重要性。**表达则是解决问题**，把发现的端倪转化为下一次优化的方向，用特征解决是把“人工”的智能加进去，试图让模型学到。其实可以借特征表达，也可以借助模型来学习，甚至手工后处理。

很多时候，发现问题比解决问题更关键。每个比赛任务的关键点都不一样，需要摸索和发现，需要实验。当然也有时候，根本没有关键，只需要心细。我，什么都不重要，运气永远是最长的那块板。

特征工程的实践

我们以山东赛民宿比赛为例，借此实践一下怎么找到关键。水平所限，我只能讲我能看到的关键。

第一步：明确数据中所含的实体与关系

这道题就是：个人，民宿，平台。推荐场景就是用户、商品、场景，制造业就是各个产线上的设备流。

明确了涉及到的实体(entity)->之后，着手开始建立实体之间的关系，建立一个商业模型。这个任务，一个人想住民宿、在平台预定、民宿确认、最终入住、离开。读完题目和具体数据后，可以发现，还有人会中途取消订单。



关系：由于民宿一个屋子同一天只能入住一个订单，所以整体订单有时间关系。房地产最重要的：位置，所以不同的民宿酒店之间存在空间关系。可能有些人住了不同的酒店，可能存在图关系，但几乎可以忽略。

确定实体关系的时候，可以确定一下训练集和测试集之间的关系。尤其是这题有时间关系。

再以几年前eleme的一道风控场景面试题为例，在外卖中检测优惠券的撸羊毛行为。这个场景就需要用户登陆APP，选择商品，支付几个环节。那么整个链条就可以从用户、APP、商品、支付等角度着手建立特征。用户特征可以包括他的财务画像、点外卖的频率、最近收货地址的个数，支付特征可以包括支付时间段、支付买家、支付金额相关。

第二步：根据商业模型和数据进行基础特征提取

确定商业模型各个环节可以提取的特征，例如根据常见的类别或数量列提取，或根据业务进行提取。这个任务就是：

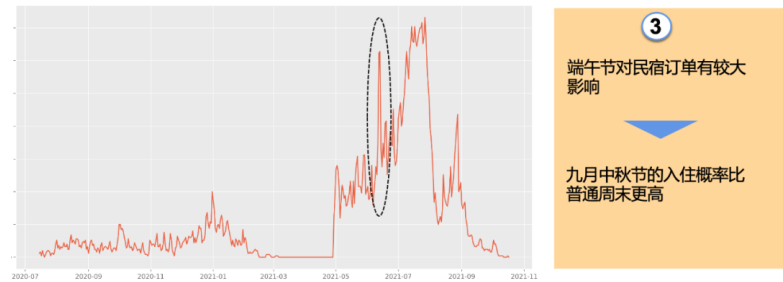
- 民宿：历史入住率，房间个数
- 个人：没有个人特征，都是无情工具人
- 时间关系：根据周期性，历史上的入住人数、入住率，节假日影响
- 空间关系：周围民宿的入住人数、入住率

我觉得亲身体验一下有助于理解业务。恰好工地不忙，就想去一下大的城市，体验一下民宿。买了票之后疫情爆发了，票今天就过期，这个主办方给报销吗？

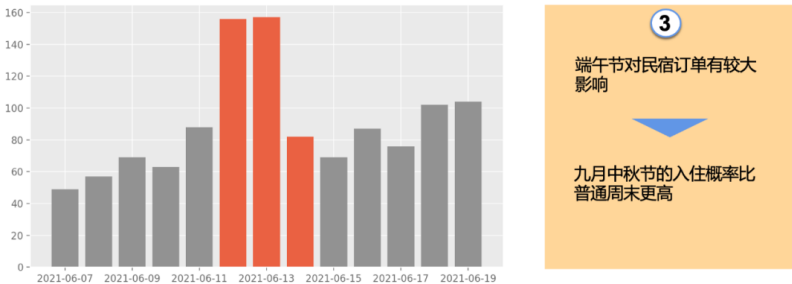


起手的道理很简单，民宿空置预测，可以用这个民宿历史空置概率作为base。

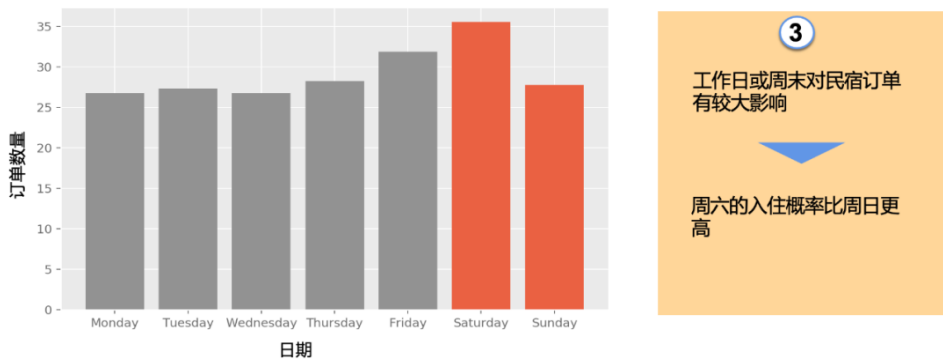
其次，是节日的影响。那么如何量化这个影响，用上一个端午节。



节日可以进一步优化。想了一下，节日三天是等价的吗？我们知道放假三天，一般第三天就要恋恋不舍地离开。所以假期最后一天和之前几天是不一样的。红色就是三天的端午节影响。



其次，周六和周日可能也不同。同理，如果是来周末度假的，可能周日就得走了。所以用历史上周六日差异来优化



都是基本操作，作为只会调参的工地老哥，感觉这次并没有发挥出我手工调参的优势。

第三步：根据数据洞察、业务敏感、重要特征进一步优化

进一步的优化，没有普遍的套路。如果对这个业务恰好比较熟悉，就可以业务先行；如果涉及到的数据比较大对业务本身又不熟，可以从解读模型重要性入手，所以因人因任务而异。

本次比赛，我尚未学习和实践到这么高深的步骤。等我学会了，



kaggle竞赛宝典

数据竞赛Top方案，竞赛黑科技，竞赛到入职的一些感想。

272篇原创内容

公众号

喜欢此内容的人还喜欢

一场偶遇，秒懂山东人为啥这么重视编制了水

为什么蜷缩起来会感觉更暖和一些潮汐朝夕

善良可爱的山东人
懂懂分享