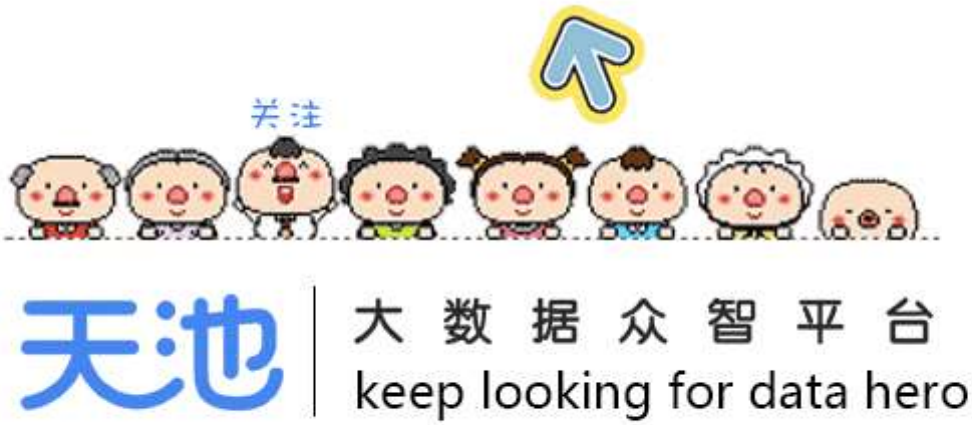
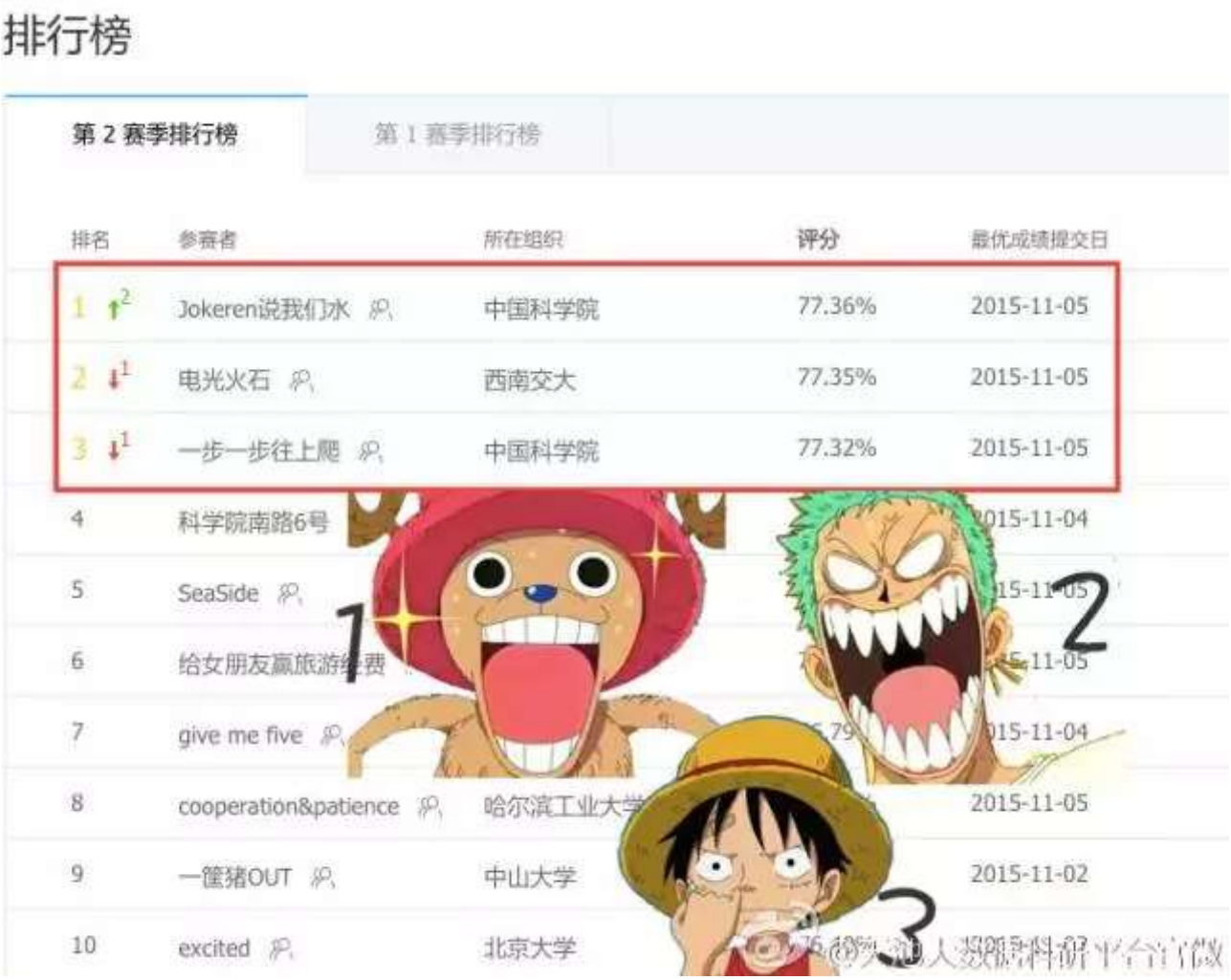
了不起的特技｜截榜之日反超冠军

2015-11-06 李强 天池大数据科研平台



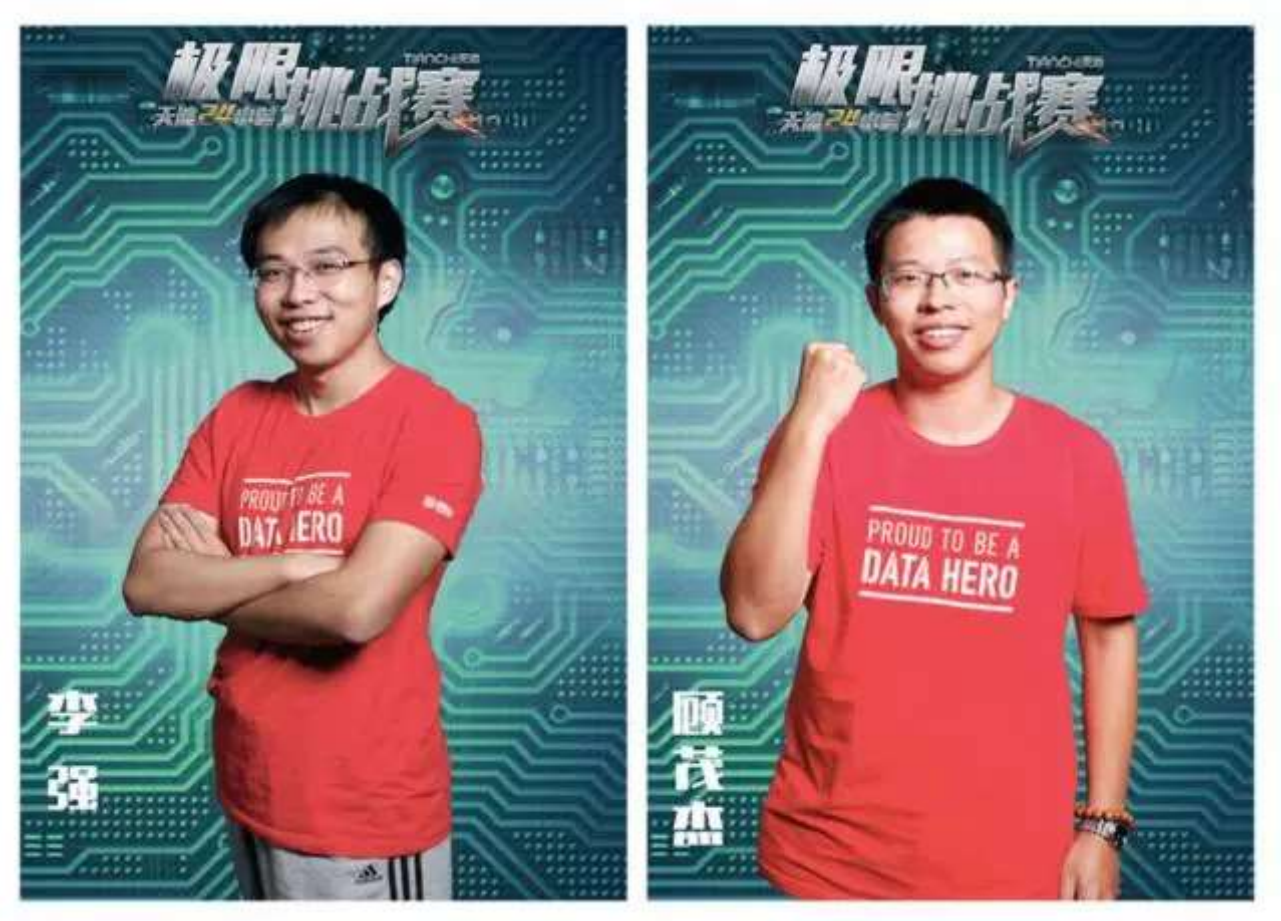
新浪微博互动预测大赛已于昨日截榜。冠军居然不是盘踞排行榜榜首已久的“电光火石”？而是来自中科院的“Jokeren说我们水”。在最后一天以0.01%的差距反超大牛，他们到底加了什么样的特技？



（新浪微博互动预测大赛截榜之日的排行，前三的同学们，这些表情有生动体现你们的心情吗）

文 | 李强

大家好，我们是此次参加新浪微博预测大赛的的“Jokeren说我们水”队，大岑神，大茂神以及我都来自中科院计算所，在研一的时候我们就是舍友。



（李强&顾茂杰参加24小时大赛的宣传照）

说到组队，有很多好玩儿的事儿，我和大茂神结对参加了LBS比赛，也算是半个老队员了，对odps和pai挺熟悉的（坑踩的挺多的），看到微博这个比赛的时候就毫不犹豫的参加了。而大岑神一直是我们的榜样，码神，acm大神，颜值爆表，下面附图。

岑武斌在鞍山赛区颁奖仪式上领取金奖（领奖者右4）

如此大腿，怎能不抱！

在lbs的时候，我就邀请大岑神组队，结果被残忍拒绝了三次，分别是头天晚上（我考虑一下），第二天中午（我问问同学），第二天晚上（我已经组队了，我们下次再合作吧！）。

被大岑神如此拒绝，让我怀恨在心。这次微博比赛初赛完结之后，考虑到我和茂杰需要去参加24小时挑战赛，微博比赛复赛最开始有很久的真空期，急需一名队员。而大岑神也想找个比赛练练手，于是我们的队伍诞生了，还缺少一个队名。为了取得好名次，我们膜拜了一下我们所的carry大神Jokeren，祈求可人大神保佑我们。

传奇的队伍就这么诞生了：
我脑洞比较大，同时是研究社交网络理论的，对模型以及融合比较了解，适合数据分析和模型训练。
大岑神码力惊人，手法很稳，适合提特征以及框架搭建。
大茂神NLP专业，基础杂实，可主攻微博文本分析。

然而剧情并不是这么发展的。。。

首先复赛刚开始，我回家玩耍，大岑神去打acm了，茂杰兄去苏梅岛拍婚纱了（此处脑补一万字）。十一结束，我跟茂杰去打24h挑战赛了。说好的十一我带着大岑神熟悉odps准备上手呢。。。。结果我给我们计算所的另外两个队普及了odps和mr的使用，于是乎他俩在我们队还没开始做的时候已经稳居前五，参加第三和第四的队伍。

等到我们十月十五开始全力投入的时候，问题接踵而至。下面是片段：

片段一
大岑神：怎么办，大强，天佑（第四）和岳志磊（第三）他们已经前三了，我们还没跑。
我：没事，他们初赛成绩没我们好，我们把初赛特征先实现了再说。
茂杰：对的，我们初赛特征应该不错，只是初赛样本少，发挥不出才被规则吊打，现在应该吊打规则吧。
第二天特征实现完毕，我兴奋的训练了熟悉的gbdt。
我：卧槽，怎么只有45%，不科学啊
大岑神：怎么办，大强，他们已经77了
茂杰：一定是姿势不对，我们讨论一下。

片段二
我：不可能，随机森林不可能效果好，好了我直播吃（哔——）
吃饭完跑完了实验，成绩65%，大家男默女泪。

片段三
我：卧槽，我刚刚训练了一个线下80.04%的模型，我们要拿到周星星了！
众人：卧槽
然后再跑了一个，发现成绩又的很差。
大家在一起排查了三天，发现是我误操作了，我被深深的鄙视了，从此我这边有了提升都需要被大家检验一遍。。。

比赛的过程很不顺利，微博内容太短，无法很好的文本分析，茂杰的优势发挥不出。我这边训练的模型又在接二连三的报喜（成绩被规则吊打），大岑神不断地给我一些特征，加进去都是然并卵。不断地迭代，加特征训练模型，线下成绩76.8%。

76.8这是一个神奇的数字，一个幸运数字，一个看起来让我们毛骨悚然的数字。这个线下成绩陪我们从特征版本一走到特征版本五。到了最后，我们已经要发狂了，加了500维特征了，成绩还这么稳定，76.8纹丝不动。唯一的提升就是修复了一个bug，线下线上成绩一致，数据分析的结果被验证了。这个过程真的是很绝望，努力很多，没有一丝提升。

一直到上周六的一个下午，所有队员在一个空会议室里面结对编程，我决定删特征，大岑神决定重头实现一遍所有特征找bug，茂杰决定调参试试，三管齐下，很激动的一下午，发现删了特征竟然成绩提升了0.1，第一次出现了76.9这个数字，好令人振奋啊。两周来唯一一次打败了76.8的魔咒，但这点提升是远远不够的。吃完饭大岑神建议去健身锻炼一下，我选择去拉拉单双杠思考人生。走的路上一个想法油然而生，我们一直被模型限制着，为什么我们不多自由发挥一下，xxx呢。

当晚实现了这个想法，第二天下午，一看，线下75.5！再被怀疑误操作，验证了半天之后，发现竟然是真的。当时的内心真的是崩溃的，半个月了，终于有了大提升。当天晚上，我调节了一下参数，深夜3点钟给出线下成绩，77.9！忐忑不安的提交了线上成绩，终于在11月2号杀入了前三。

最后三天终于能看到了胜利的希望，连续通宵了几晚，疯狂的实现想法，线下成绩稳步提升，最终在最后一天，我们逆袭了第一。

干货来了！！！

说了这么多，说一下干货吧：

1、 这次比赛第二赛季是多分类问题，跟第一赛季差异很大，不要被经验误导。第一赛季效果不好的方法，第二赛季不一定没用。所有第一赛季尝试过的方法，重新尝试一遍吧，会有奇效的。

2、 不要被以往的经验误导，经验告诉我们，GBDT吊打RF吊打LR，然而这个微博比赛呢，大家都懂的。还是尝试，只有多尝试，才有发言权说哪个好用哪个不起作用。

3、 LDA的使用，首先建议看一下LDA的原理，简要了解一下（我是没完全看懂。。）。然后再看一遍LDA的实战贴！纸上得来终觉浅，原理贴是不会告诉你去过滤停用词的（组里有人懂nlp的好处），原理贴也不会告诉你应该选多少个主题的！

4、 数据分析的重要性。拿到了数据，不要立即想着提特征，实现算法。而要大体知道数据是什么样的，数据的分布如何？统计一下微博等级的分布，观察一下数据，去读一读微博的那张表。做好了前戏后面才能更快的上手，有了数据分析的铺垫才能取得更大的进步。

5、 还是数据分析，看一看模型的输出结果，究竟是哪些微博被误判了，是屌丝被误判成高富帅的多，还是高富帅被误判成屌丝了，接下来才知道应该提取什么特征，还有哪些不足。

6、 定期检查！定时检查一下现有的代码，从头梳理一遍，看看有没有bug，有bug的特征不如不加。画出特征思维导图，在画的过程中，你就会发现，哦，我好想这个特征还没加，这俩特征是不是可以叉乘一下。

7、 快速的代码迭代，不断地去验证想法。有的时候，你有了一个想法，不知道好不好用，就先实现了，然后结合以前的结果看看效果。不断地往特征池里面添加会使最终结果变好的特征

8、 最重要的，团结一心，永不放弃。这次比赛我们被吊打全程，中间也一直卡在76.8无法提升，不断地迭代特征，也没有任何提升，这真的是很崩溃的。团队的信念和士气就很重要，团队对于胜利的渴望让我们在最后几天的冲刺阶段连续通宵依然斗志满满。

最后一天夺得了第一，我们是兴奋之余心有余悸。兴奋的是功夫不负有心人，我们终于登顶了。同时也在担心，中间的任何一步，一旦失误了，我们可能就与第一失之交臂了。总的来说，这次比赛收获很大，也很刺激。也知道了，比赛要提前开始打，最后的时刻我们最希望的就是能再给几天，多几次提交机会。

长按二维码关注天池小报，让我走进你心里。