139 | 数据科学团队怎么选择产品和项目?

2018-08-22 洪亮劼

AI技术内参

进入课程 >

讲述：初明明

时长 07:24　大小 3.39M

上一期内容，我们聊了聊数据科学团队在工程流程方面所需要注意的三大问题，分别是代码管理流程，开发部署环境流程和数据管理流程。

今天，我们来继续讨论数据科学团队发展这个话题，来看另外两个关键问题：如何选择合作产品以及如何选择项目。

如何选择合作产品

选择什么样的产品进行合作，对于一个数据科学或者人工智能团队的发展来说，是一个非常重要的问题，是决定工作能否事半功倍的关键步骤。

作为工程技术团队，很多数据科学或者人工智能团队都需要支持多个产品，或者说是有机会选择产品。一个稳定的产品往往可以让一个人工智能团队得到快速健康的发展，并且能够逐渐形成良性循环，发展到可以支持更多的产品。

那么，什么样的产品是值得合作的产品呢？

我先来说一类需要谨慎合作的产品，那就是全新的产品方向。对于全新的产品来说，公司之前在这个方向没有太多的产品积累，也可能完全没有技术积累。对于这一类项目，我们需要格外小心，特别是当你的团队还在发展的初期。

新产品有一个特点，那就是极大的**不确定性**。产品范围、需求和时间一般都是不确定的，这些都是一个稳定项目的天敌。另外，还有一个比较棘手的问题，那就是新的产品方向，特别是公司以前从来没有研发过的项目，往往**缺乏数据**。对于机器学习来说，数据匮乏就是“巧妇难为无米之炊”了。

然而从另外一个角度来说，新的产品方向往往又能得到公司高层的重视。毕竟新产品往往是公司“新的赌注”（New Bet），所以很多时候也能够得到不少团队的支持和资源的倾斜。

那么，在这种情况下，如果你是数据科学团队的负责人，你就要对这个新产品的利弊有一个充分的认识。如果你的团队已经相对比较成熟，有好几个稳定的产品支持，还有一些剩余的资源可以分配，那么接受一个全新的产品也不失为一种尝试，虽然有一定的风险。最坏的情况是这个产品方向完全失败，但是不会对团队产生致命的影响。

了解了这种风险比较高的产品之后，我们来看一看什么类型的产品更值得一个人工智能团队来合作。总结来说，这样的产品一般需要满足以下两个方面。

第一，看方向，这个产品最终是需要数据来驱动的。如果一个产品最终会产生大量的数据，而且这些数据能够表征这个产品方方面面的发展，那么，针对这样的项目，人工智能可以起到巨大的推动作用。

第二，看地位，这个产品是公司的核心发展项目。这一点看似容易识别，但是在一个相对比较大的公司里，有时候反而并不那么容易识别。那怎么判断呢？介绍一个相对比较简单的方法，就是看一个产品是否和公司的利润或者说是公司的核心用户数据有关系。因为从公司的

层面看，一个团队的投资回报率很关键，很多时候会以此来决定是否继续支持这个团队的发展。

举个例子，有这么一个产品，虽然不直接产生公司的利润，但是能够帮助公司增长用户，那这个项目也可以算是公司的核心项目。**能够支持公司的核心项目，是人工智能团队稳定快速成长的基石。**

如何选择项目

选择好了合作的产品之后，一个产品的迭代过程中还会产生很多不同的项目。是不是这些项目都值得做呢？接下来我们就来看一看究竟应该选择什么样的项目来做。

关于项目的选择，我们先有这样一个共识。在团队和产品发展的不同时期，对于如何选择项目应该有不同的考虑。

在这里，有一种思维模式可以帮助我们来对不同的项目进行筛选，那就是“**投资组合**”（Portfolio）的思维。通俗地讲，投资组合思维的核心就是“不能把所有的鸡蛋都放在一个篮子里”。

简单来说，我们可以利用四象限法，把不同的项目分为“高投入、高回报”、“低投入、高回报”、“高投入、低回报”和“低投入、低回报”这四种类型。那么，针对这四种不同的类型，我们就需要利用投资组合的思路来选择项目。

从理论上来说，我们希望所有的项目都是“低投入、高回报”的，肯定不希望项目是“高投入、低回报”。那么，从表面上看，对于项目的选择，我们其实并没有太多可以争议的地方。但是，实际的情况是，对于绝大多数项目来说，我们并不知道这个项目属于什么类型，最好的情况也无非是对项目有一个估计，而这种估计很有可能会和真实情况相差甚远。

因此，对于投资组合的思路来说，这不仅仅是一种对于投入和回报的估计，还包括对于不同类型项目的选择。这里呢，我就讲一些在以往工作中积累的项目选择经验。

一般来说，**对于一个人工智能项目来说，特征工程（Feature Engineering）都是属于“高回报”的项目。**对于大多数的类似项目来说，特征工程往往能够针对项目带来本质上的提升。寻找到好的特征是一个项目能够持续成功的重要途径。

在“高回报”的情况下，我们需要考虑的是，这个项目是“高投入”的？还是“低投入”的？如何评价一个特征工程项目的投入成本呢？我们问以下三个问题。这个项目可以基于现在的数据链路（Pipeline）来做吗？是否只是计算一些数据的简单统计量？是否只是把每天不同的统计量做一些叠加？如果三个问题的答案都是“是”，那么这种类型的特征工程项目就属于“低投入”。

在一个产品的早期，应该尽量尝试这样的项目，快速发现有用的，特别是那些能够让产品的效果得到迅速提升的特性。而且，因为特征工程“高回报”的特点，在产品迭代的任何一个时期，我们其实都可以关注某一部分的特征工程项目。只是说，也许在产品的初期，找到一系列特征，或者说挖掘出一系列有效的特征，往往会非常容易；但是在产品的中后期，难度就要大一些。

我介绍的另一个经验可能和你想象的不太一样，**对于核心算法（例如搜索、推荐、广告）的改进，比如改进排序算法属于“高投入”的项目。**这类改进算法项目的一个明显特点，就是往往需要有较长时间的研发周期。而且，这类项目的升级换代往往需要“基础设施”（Infrastructure）或者平台级别的变化。也就是说，这类项目的投入比较大，周期比较长。

当然了，这类项目的回报如何，其实并不是特别容易估算的。例如，我们在特征不变的情况下，从线性模型更改到树模型，可能会有 5%~10% 的性能提升。但是更改到树模型之后，也许我们还能够加入更多适应于树模型的特性，带来后面的 10%~20% 的提升。因此，如果考虑到一个回报的系列性效果，算法的更改升级还是应该引起我们足够的重视的。

小结

今天我为你讲了数据科学团队的两个核心问题，那就是如何选择产品以及如何选择项目。

一起来回顾下要点：第一，我们聊了聊产品的选择，尽量对新产品持谨慎态度，同时尽量支持并开发公司的核心产品；第二，我们分析了如何选择项目，重点是对四种类型的项目进行一个探讨。

最后，给你留一个思考题，如果我们希望从树模型升级到“深度模型”，这种项目属于我们介绍的四种类型项目中的哪一类呢？

欢迎你给我留言，和我一起讨论。

上一篇　138 | 数据科学团队必备的工程流程三部曲

下一篇　140 | 什么是计算机视觉?

精选留言 (2)

写留言

阿国
2018-08-26

高投高产

展开

阿国
2018-08-26

高投高产的

展开