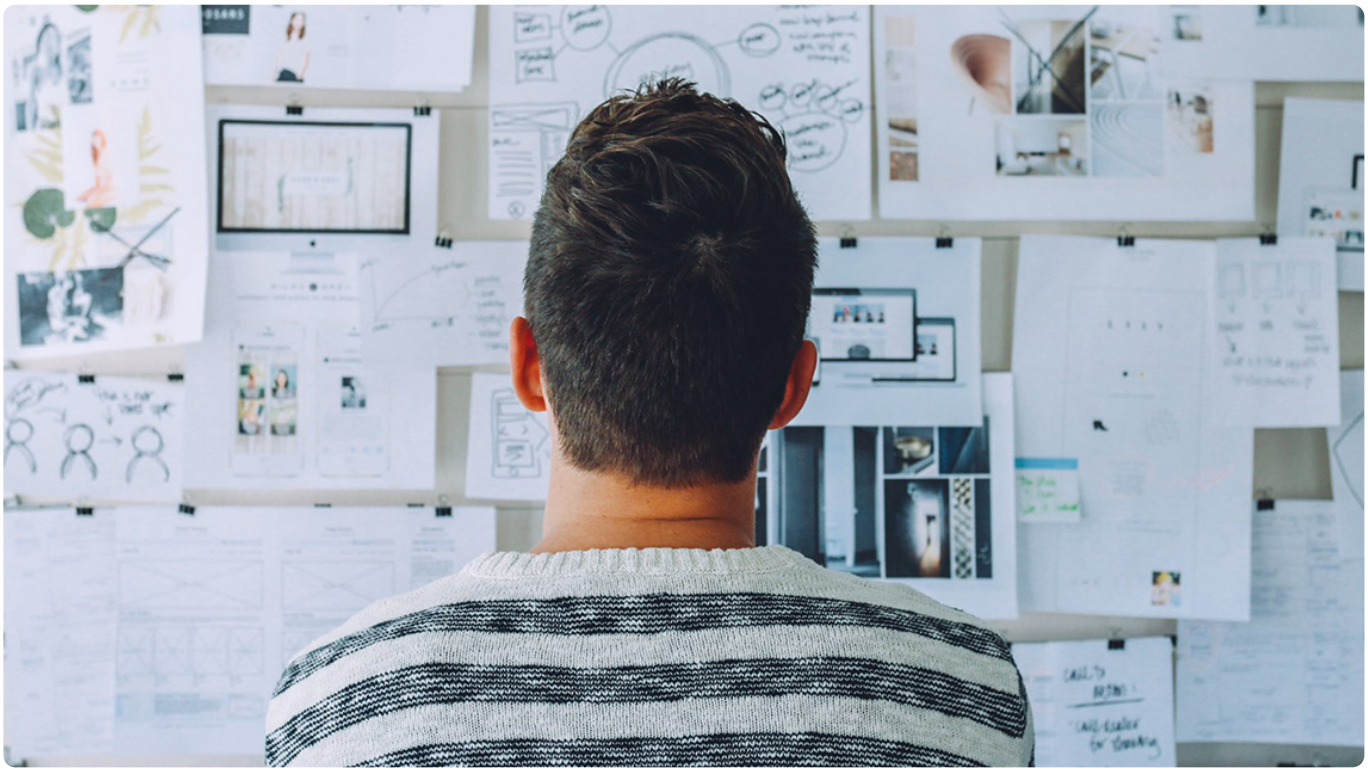
001 | 如何组建一个数据科学团队？

2017-10-10 洪亮劼

AI技术内参 进入课程 >



讲述：初明明
时长 11:17 大小 5.17M

数据科学团队眼下已经成了很多数据驱动型公司的标准配置，数据科学家也成了最“性感”的职业。不少公司都在想办法建立或扩展自己的数据科学团队，而究竟需要什么样的数据科学团队，成了很多公司在发展过程中都会遇到的棘手问题。

在目前的职业市场上，有各种背景、各种经历的人都自称为“数据科学家”。那么，如何从这个蓬勃发展，却鱼龙混杂的人才市场中找到合适的团队成员呢？今天我就来和你聊一聊作为一个工程团队的负责人，或者一家公司的 CEO，该如何招聘并打造自己的数据科学团队。

数据分析还是算法模型

目前人才市场上大致有两类数据科学家，一类偏数据分析，一类偏算法模型。因为这两类人才的区别，不同公司乃至同一公司的不同数据科学团队就有了差别。在招聘之前你必须明白，这两类数据科学家的特质很难在同一个人身上体现出来。也就是说，你必须根据当前公司和团队的需求，来决定目前应该招聘更偏重数据分析，还是偏重算法模型的数据科学家。

先来说说偏重数据分析的数据科学家，他们可能来自于统计、数据分析等学科，也可能来自于很多需要数据分析的自然科学学科，比如实验物理、生物、计算化学等。作为团队的负责人，你需要重点考察候选人是否系统学习过数据分析的相关课程，是否具备数据分析的基本能力。下面我从理论知识和实际应用操作两个角度来和你介绍下考察要点。

从理论知识的角度来说，你需要考察候选人：

1. 是否对概率统计有基本的认知；
2. 是否能够使用基本的假设检验对数据进行分析；
3. 是否对高级的假设检验方法，比如非参数假设检验（Nonparametric Hypothesis Testing）有所了解，能否快速学习和查询到相关的方法；
4. 是否了解 A/B 实验，并基本了解实验设计；
5. 是否了解高级的因果推论（Causal Inference）工具，并能够使用简单的因果推论工具对实验数据进行分析；
6. 是否了解如何对时间序列下的数据进行合理分析。

当然这些技能只是作为数据分析候选人的一些基本素质，具体还需要和领域知识相结合。

从实际应用操作的角度来说，你还需要考察候选人：

1. 是否熟悉一些基本的数据分析工具语言，比如 R 或者 Python；
2. 是否对 SQL 有所了解；
3. 是否对 Hadoop 等大数据处理工具有所涉猎；
4. 是否了解一些基本的计算机算法。

同样的，这些也是基础素质，还需要和具体的职位相结合，你才能考察候选人的综合情况。

接着我们来看偏重算法模型的数据科学家，他们主要来自于计算机科学、计算机工程、电气工程等工程方向。你需要重点考虑他们是否有基本的算法建模能力；是否系统地学习过算法、机器学习、统计分析等课程；是否在实际工作中，有系统的相关开发经历；对数据的认

识，特别是对数据驱动型产品是否有基本的了解。下面我依次从理论知识和实际应用操作两个角度来谈谈具体的考察内容。

从理论知识的角度来说，你需要考察候选人：

1. 是否对概率统计有基本的认识；
2. 是否对传统的机器学习算法模型有所了解，包括分类、回归、聚类等；
3. 是否对概率图模型有所了解；
4. 是否对深度学习模型有所了解；
5. 是否对优化算法有所了解；
6. 是否有基本的计算机算法、数据结构、数据库、操作系统的知识；
7. 是否对某一些特定领域内的模型有所了解，包括但不限于信息检索、推荐系统、计算广告系统、计算机视觉、文本挖掘和分析、自然语言处理。

这些，特别是第 1 项到第 6 项是候选人的基本素质。第 7 项是针对某一个具体的职位所需要的背景知识。

从实际应用操作角度来说，你需要考察候选人：

1. 是否可以使用某种计算机语言（比如 Python、C++、Java、Scala）来实现一些机器学习算法；
2. 是否可以使用和扩展现有的机器学习工具（比如 Scikit Learn、XGBoost、Vowpal Wabbit 等）；
3. 是否可以使用以 Hadoop 为基础的大数据工具（比如 Hive、Pig、Spark 等）来构建生产环境；
4. 是否对深度学习框架（比如 TensorFlow、Caffe、MxNet、Torch 等）有所了解。

这里列出的也是一些基础素质，还需要和具体的职位相结合，来考察候选人的综合情况。

总体说来，如果你希望招聘的职位更偏重于理解现有数据，通过数据来对公司或团队的下一步决策有所帮助，那么这个职位就更偏向于数据分析。如果你希望通过算法和模型来改进你的产品，无疑你需要招聘一个算法模型方向的数据科学家。

小团队、大团队

不同的团队往往需要不同的数据科学家配置。即便是同一团队，在不同时期其实也需要不太一样的设置。我这里讲的是一些基本的团队设置理念，具体的团队还需要根据不同的领域有所调整。

总体说来，在团队比较小的时候，甚至是初创公司的团队，你需要具有“通才”性质的数据科学家，而在团队扩大、公司稳定以后，你需要各类“专才”性质的数据科学家。

团队比较小的时候，我们可能只需要招聘一两位数据科学家。这个时候的数据科学家必须同时承担数据分析和算法建模两个角色。有些情况下，这时候的数据科学家其实更偏向于“数据工程师”（Data Engineer）的角色，那就是和其他工程师一起搭建公司的数据平台，对数据的引入、整合、清理提供支持。

早期时候，因为公司内部基础设施的限制，数据科学家往往需要花费大部分时间在数据平台和通路的构建上。这时候，其实很难形成有效的数据分析和算法建模工作。从另外一个角度来说，在公司非常早期，也就是在数据平台还没有一个基本雏形的时候，招聘和建立数据科学团队是不现实的。当有了基础的数据平台时，和数据有关的工作一般就是计算一些简单的产品运行指标（Metrics）然后在仪表盘（Dashboard）里展现出来。能够达到这一阶段后，一个团队或者公司才具备了建立数据科学团队的最基本条件。

小团队所需要的“通才”数据科学家有两个内涵。第一，在初期，数据分析和算法建模都同样重要。甚至在有些情况下，数据分析有着更加急迫的需求（因为需要人为地对产品迭代进行决策）。这个时候，以数据分析为主导的数据科学家要能够对现在的产品需求有很强的理解，能和产品经理、其他工程师一起快速分析产品的问题，为产品迭代的决策提供数据支持。

第二，在初期，绝大多数产品所需要的算法和模型都并不复杂，甚至仅仅需要数据科学家部署最基本、最简单的算法。因此这个时候，即便有算法建模需求，也只需要数据科学家有比较广的知识就行，能够快速识别和实现最基本的模型。在这个阶段，对某一个方向有着深厚背景的“专才”往往并不能体现出优势。

当业务逐渐稳定并且扩展以后，团队也逐渐扩张，小团队的“通才”模式就慢慢不太适应组织的发展了。这个时候，我们需要针对目前的产品和业务招聘“专才”数据科学家。一般来说，我们需要有一部分数据科学家负责数据分析方面，需要另外一部分数据科学家负责算法和模型开发方面。这时候单个人往往已经不能胜任两方面的任务了。

从数据分析的方面讲，“专才”的模式需要我们更细地区分开两类数据科学家，一类是负责设计 A/B 实验、设计和分析产品指标的专项数据科学家，另一类是对各个产品领域进行长时间分析数据内涵（Insights）的数据科学家。

从算法建模的方面讲，“专才”模式往往就是针对不同的业务流程线，需要招聘单独的人才，比如针对图像处理的人才、针对搜索系统的人才、针对推荐系统的人才。这个时候，能否招聘到称职的领域专家，成了团队和产品能否持续良性发展下去的根本因素。这个阶段招聘需要注意的问题是，不要寄希望通过招聘“通才”来发现“专才”，因为从“通才”到“专才”的训练是需要很长时间的，这里面有短时间内不可逾越的鸿沟和难以积累起来的经验。所以，当公司和团队发展到一定规模的时候，分清形式进行“专才”招聘是必须要进行的任务目标。

小结

今天我为你简单分析了如果要组建一个数据科学团队，你需要招聘什么样的数据科学家。一起回顾下内容要点：第一，偏数据分析和偏算法建模的两类数据科学家在技能背景方面有很大区别；第二，“通才”和“专才”在公司或团队的不同阶段承担着不同的角色。

最后，给你留一个思考题：如何在筛选候选人简历的时候，就能够比较准确地了解这位候选人的经历和能力更偏向数据分析还是偏向算法呢？另外，如果你想成为数据科学团队的一员，不妨对照今天我们聊的考察要点自测一下，看看接下来还需要在哪些方面继续努力，做好积累呢？

期待你给我留言，和我一起讨论！

上一篇　开篇词 | 你的360度人工智能信息助理

下一篇　002 | 聊聊2017年KDD大会的时间检验奖

精选留言 (11)

写留言

阿祺 置顶
2017-10-24
3

讲得很好，重点很多而且不冗余sweet and short，也为我今后找工作或者创业理清了不少思路。非常感谢分享！

Momo
2017-10-31
7

对不同角色的要求清晰地定义了两个job model，给想做数据科学家的同学制定了两个清晰可触达的目标

作者回复: 是这样的。

帅帅
2018-10-20

4

好尴尬，我就属于文中的通才。

一方面，能做数据分析，会使用hadoop/hive/spark做数据打点接入、处理建模、出dashbord报表，供老大和产品进行迭代决策；
...
展开

Lynn
2018-05-25

1

如何在筛选候选人简历的时候，就能够比较准确地了解这位候选人的经历和能力更偏向数据分析还是偏向算法呢？
——————————————————————————
主要看候选人的所学专业背景（统计相关or计算机相关）
技术背景（分析工具用的多还是算法模型用的多）...
展开

Drowning ...
2017-12-26

1

请问，有比较好的数据分析的课程（或其他资源）推荐么？突然发现数据分析的相关的理论知识第一次听说。

hallo128
2019-01-21

1. 项目经历：
数据分析——统计假设检验建模
算法——了解基本算法，深入理解各个算法的细节，能够提出改进思路
2. 知识贮备
数据分析——传统统计课程：假设检验，试验设计，抽样调查...
展开

杰之7

2019-01-19

通过这一节的阅读和结合我自己的看法，我觉得如果候选人能在一定时间能通过各项运营指标发现其中的问题和落地改善方案，这种候选人偏向数据分析。如果后选人能通过不断优化产品的性能指标，推出更快速准确的数据模型，这种候选人更适合算法类。

目前我还没有正式接触过数据科学的工作，也不是科班生，基本素质和职位背景都需要...

展开

陈星强•De...

2018-10-01

很不错👍团队小 就得啥都干。干好了就行，团队大，专人专事，一人一责

展开

蒋鑫

2018-09-12

数据分析，更偏向产品数据的指标性分析，帮助达到业务目标，往往与产品和运营人员配合；算法建模，更偏向产品某个模块本身，如何挖掘数据背后的规律和价值，提升产品功能，与产品人员深度合作。

展开

一只豆

2018-04-29

因为交叉的职业经历，我是有领域经验的互联网产品经理，也是公司创始人。一方面需要搭建团队，另一方面希望能够成为完美配合数据科学团队的领域专家。关于前者，老师已经说得足够清楚，现在想请教的是关于后者：完美配合数据团队的领域专家的内功心法是什么？场景需求的抽象化？还是什么别的？

Steve 大...

2018-01-02

「对各个产品领域进行长时间分析数据内涵（Insights）」，这句话不太理解，您能解释一下吗，谢谢