<mark>算法工程师的术与道:从特征工程谈数据敏感性</mark>

原创 包包闭关修炼 包包算法笔记 2020-12-11 11:50

收录于话题

#算法成长之路

目录

问题导入 什么是数据敏感性 如何衡量数据敏感性 数据敏感性如何培养 导入问题的答案

面向人群: 算法工程师

问题引入

我们从一道题目说起,

请不择手段地提高下面这个数据的label预测精度。

特征输入: id和id_sub标识数据, question是问题, answer回答

标签: lable是否匹配。

数据展示:

label	answer	question	id_sub	id
0	不一样	采光怎么样	0	5631
1	东向一上午采光,最多四个小时	采光怎么样	1	5631
1	西向一下午采光,最多6个小时	采光怎么样	2	5631
1	西向阳光最好的时候,12点就开始进光了	采光怎么样	3	5631
1	一直到下午5点多	采光怎么样	4	5631
0	有的	有房子看吗	0	5632
0	刚刚那套房子业主自住的	有房子看吗	1	5632
0	这套也是业主自住的	有房子看吗	2	5632
1	采光挺好的 楼层高	采光怎么样	0	5633
0	这样白 你先看看这个户型	采光怎么样	1	5633
0	主要也是因为大部分房源出租了 看房确实不太随时	采光怎么样	2	5633
0	你这边现在住在哪啊?	采光怎么样	3	5633
0	是啊	采光怎么样	4	5633
0	您好,您之前咨询的房源户型很优质,您有时间去看看吗?	这个小区对应哪个幼儿园和小学	0	5634
0	房子看好了吗?	这个小区对应哪个幼儿园和小学	1	5634
0	您好,您之前咨询的青山溪语小区有新上房源,不知道您是否感兴趣?	这个小区对应哪个幼儿园和小学	2	5634
1	您什么时候方便	可以看吗?	0	5635
1	我约一下	可以看吗?	1	5635
對態	二套税妻37万多点	可以看吗?	2	5635

算法工程师的术与道: 从特征工程谈数据敏感性

问题答案匹配预测任务

黄金玩家:这个简单,BERT跑几个单模型融合一下,这玩意不是全靠ensemble吗?

王者玩家 (Kaggle GrandMaster) : 数据中有非常强的规律,我可以用特征工程来刻画出来。

想看这个题目的答案,直接翻到文章最后。

什么是数据敏感性

下面我们引入本文的正题,工业界的算法工程师的核心竞争力到底是什么?

这里引用认识的算法大佬的一段观点。大家比较认同的基本功是编程,数学,数据洞察。随着模型工具化,实际工作中对数学的要求也会 比较低,在编程方面,算法工程师明显没有比系统开发更有优势。那么我们不仅想问,算法工程师的核心竞争力到底是什么?我们用排除 法,最后剩下一项叫数据洞察,也叫数据敏感性。

数据敏感性听起来很玄乎。每个数据分析,商业分析,产品经理的岗位的JD都写了要求数据敏感性要高?那么这个东西怎么量化呢?到 底是刻画一种怎么样的能力呢?如果用一句话来说,就是发现数据规律的能力。

这个具体在不同的岗位上有不同的体现,比如算法工程师的特征工程水平高低。这里比较认同一点,数据encoding只是基本的特征工 程,而业务理解的数值转化,数据规律的特征刻画是高阶的特征工程。

在商业分析岗位,其体现是给你一组数据,快速发现其中的存在的异常和问题,并指出可能的原因。

对数据分析岗位来说可能就是给你一些数据指标,在找出其中的相互关系。

如果用一句话来说,就是发现数据规律的能力,化繁为简,范式地表达数据。后者是从数据中提取信息,加工成知识的一种能力。

如何衡量数据敏感性

如果你觉自己的数据敏感性还不错,请回到开头回答那个问题。

培养数据敏感性

我们承认天赋但不谈天赋,我们看一下后天切实可行的可以培养提高的方法。

1. 掌握基本的数据分析方法,各种指标的计算方法。

大概就是这个指标是什么意思,为什么提出这个指标,这个指标能刻画什么问题。这个指标正常的表现是什么样子。

2. 理解数据背后生产逻辑和运行逻辑。

比如从广告投放到点击到下单,这个漏斗模型中各个数据是怎么得到的,他们的关联性是什么,影响因素是如何影响数据运行的。

3. 快速归因, 定位问题。

这一阶段进入分析这一件事情,之前都是在理解数据,分析数据也是为了发现问题,找出问题,解决问题。一些基本的方法论,比如控制 变量法等等。

4. 理解业务特性问题,发现本质,解决问题。

做以上的目的还是为了How出发点的。也就是尝试做人肉决策。一个最典型的能力是,给你一个任务,你能否拆分清楚,并且提炼出其 中对最终收益最大的部分。

当然以上都是数据敏感性的道,具体还有一些术。

比如:

- 客群拆分
- 分维观察
- 分布监测
- 采样观察
- 特征工程
- AB测试
- 归因分析

等等

这个就有点类似于算法工程师的特征工程实操了,比较零散,本质我们还是用手里已有的工具,去发现,刻画,解决问题。

答案揭晓

最后公布一下开头问题的答案吧

数据中的道:标签中存在连续的1,因为构造数据没有shuffle,连续的回答question的answser紧靠在一起,并且都是同一个answer。 所以这不仅仅是个NLP的问题,当然这也是聊天中存在的问题,就是有人习惯一个问题,分好几句来回答。

数据中的术:特征工程,特征一,交叉验证得到OOF概率上下错位(即时间序列中的lag特征),特征二,OOF概率的一阶差分。等等还 有不少特征可以刻画。

名词解释 OOF

有人说(玩比赛)这东西没用(leak,业务用不上),我说他有用(锻炼数据敏感性),这是数据洞察,资深算法中的数据洞察讲究发现 规律。他说非要试试,上来就是一个反转二叉树,一个链表找环,我全都防出去了。他突然红黑树袭击我脸,我大意了,没有闪。我 劝各位面试官耗子尾汁,不要老欺负玩比赛的老同志,面试要以发现候选人的闪光点,要讲武德。算法行业要以和为贵,谢谢各位面 试官老师!

关于文中数据敏感性的部分,只写了一些个人的见解,后面学习思考一下继续补充完整。

文章推荐

"

历史精彩文章:

工业界文本分类避坑指南

从一道数学题面试题到GBDT原理的推导

Kaggle GM qrfaction:数据竞赛方法论看这一篇就够了

Kaggle进阶:显著提分trick之指标优化

一文梳理文本分类技术脉络



喜欢此内容的人还喜欢

可能是全网特征工程实操最通透的...

包包算法笔记