数据分析

2021年3月29日 13:28

Excel （略）

数据透视表，vlookup，常用函数的使用，基础图表的制作

SQL

https://zhuanlan.zhihu.com/p/38354000

1. **简单查询**

- %表示任意字符串
- 利用count()统计数量

2. **汇总分析**

- 利用sum()计算总和
- DISTINCT 用于返回唯一不同的值。
- 分组group by
- 对分组结果指定条件 having
- 排序order by asc升序 desc降序

3. **复杂查询**

- 子查询
- limit 从查询结果中取出指定行，例如前n个，top(n)，排序后limit
- in 和 not in

4. **多表查询**

联结

select * from kemu left join score **on** kemu.id = score.id

- left join 就是“左连接”，表1左连接表2，以左为主，表示以表1为主，关联上表2的数据，查出来的结果显示左边的所有数据，然后右边显示的是和左边有交集部分的数据。
- right join 同理
- join，其实就是“inner join”，为了简写才写成join，两个是表示一个的，内连接，表示以两个表的交集为主，查出来是两个表有交集的部分，其余没有关联就不额外显示出来

case表达式

```
CASE WHEN <表达式> THEN <表达式>
    ELSE <表达式>
    END
```

5. **窗口函数**

```
<窗口函数> over (partition by <用于分组的列名>
                 order by <用于排序的列名>)
```

partition by用来对表分组

order by子句的功能是对分组后的结果进行排序

group by分组汇总后改变了表的行数，一行只有一个类别。而partiition by和rank函数不会减少原表中的行数

<窗口函数>的位置，可以放以下两种函数：

1） 专用窗口函数，rank（1114），dense_rank（1112），row_number（1234）等专用窗口函数。

2） 聚合函数，如sum. avg, count, max, min等
窗口函数原则上只能写在select子句中

窗口函数解决3类问题：
1）分组排名2）topN问题3）每个组内比较

练习题： https://zhuanlan.zhihu.com/p/152233908

Python

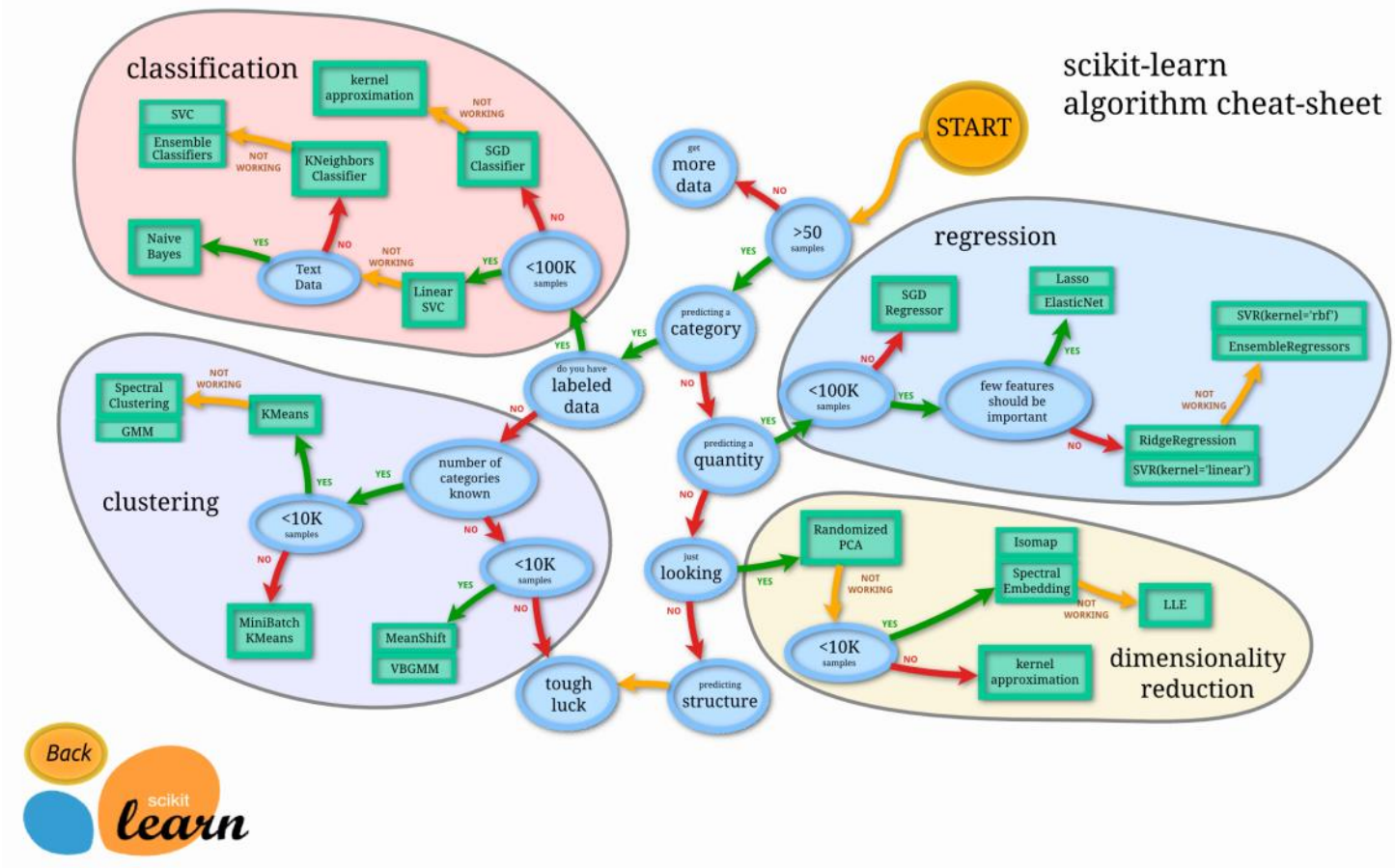
- python基本语法
- 数据分析包 numpy pandas matplotlib
- 使用python操作结构化数据，进行数据清洗，数据抽取，数据可视化
- python操作数据库

Hive、spark

统计概率（概率论与数理统计）

- 见计算机笔记中的《概率论》和《数理统计》

机器学习



分类算法：见计算机笔记中的《分类算法》
回归算法：见计算机笔记中的《回归算法》
聚类算法：见计算机笔记中的《聚类算法》

降维算法：见计算机笔记中的《降维算法》
特征工程、模型评价、交叉检验 见计算机笔记中的《其他》

业务问题

- **数据分析思维（玄学）**
  数据分析师必备的20种分析思维（都是噱头）
  https://zhuanlan.zhihu.com/p/83138160
  业务指标：信度与效度、平衡、分类、矩阵、管道/漏斗、相关、远近度、逻辑树、公式化、溯源、时间序列、循环/闭环、A/Btest、极端化、反向、队列分析、假设、归纳、演绎、指数化

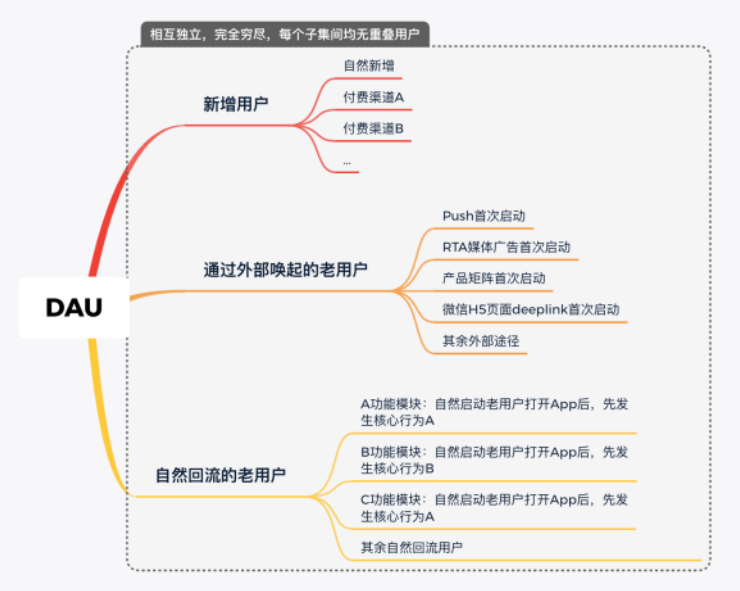
- **漏斗思维（玄学）**
  漏斗思维 https://www.jianshu.com/p/32414ca5895f
  五个经典漏斗模型，看漏斗思维穿透流程化的本质
  http://www.woshipm.com/operate/3269415.html

- **金字塔原理（玄学）**
  金字塔原理（简析） https://zhuanlan.zhihu.com/p/44423303
  什么是金字塔原理，如何运用金字塔原理？ https://www.sohu.com/a/402879711_404038

- **DAU下降**（日活跃用户数量）
  《方法论分享：DAU下降该如何分析 》https://www.sohu.com/a/449047677_114819
  一、梳理你所在公司的用户增长模式
  二、搭建数据监控预警体系
  1. 判定DAU是否异常
     常用的方法是：看日环比绝对值、周同比绝对值、日环比（就是某一天与它的前一天的数据比）、周同比、以及最近30天的变化趋势。
  2. 构建DAU拆解的指标体系
     拆解的第一层级为：
     DAU = 当日新增用户 + 首次外部唤起App的老用户 + 首次自然启动App的老用户

     拆解的第二层级为：

     新增用户。
     可按照渠道、机型等维度来进一步拆分，以细化异动是哪个拉新渠道的问题；
     外部唤起App的老用户。
     可按照首次唤起App的入口（比如Push、RTA广告、微信等）拆分。即 首次外部唤起App的老用户数= 首次push唤起App的老用户数 + 首次RTA广告唤起App的老用户数 + ...
     自然启动App的老用户，可按照用户访问App的目的进行拆分。
     具体是，通过对用户站内核心行为的先后发生顺序进行分析，将DAU的贡献归因于首个核心行为模块。比如，某产品的渗透率较高的核心模块有：A、B、C，那么自然启动App的老用户 = 站内首次进入A模块的老用户 +站内首次B模块的老用户 + 站内首次C模块的老用户 + 其他行为模块的老用户。「其他行为模块的用户」，是指未发生以上前面任何核心行为的用户。



3. 构建DAU拆解的指标体系
   指标Xi的波动贡献度=指标Xi的变化幅度/DAU的变化幅度。
   贡献度越大，说明该指标对波动的解释效果越好。通过波动贡献度，我们将异常原因定位到是新增用户、还是外部唤起老用户、还是自然回流老用户部分有问题。

实例

《数据百问系列：DAU为何会骤降？》（很有启发）
https://cloud.tencent.com/developer/article/1522874
《【数据分析】产品日活DAU下降，怎么分析》
https://blog.csdn.net/YYIverson/article/details/103844966

- **A/B test**

(分割测试或桶测试）是一种将网页或应用程序的两个版本相互比较以确定哪个版本的性能更好的方法 https://cloud.tencent.com/developer/article/1496302
小概率事件、t分布、z分布、卡方分布、p值、alpha错误、belta错误

面经问题汇总

辛普森悖论
抽样-假设前提是什么
大盘数据下降百分之三十怎么分析-拆解的原则
特征处理特征变换用到哪些办法
机器学习用哪些算法
决策树-过拟合，函数中有哪些参数
数据切片
数据可视化

女装销量下降了分析原因（拆解+定位原因）

分组求第一名，会问有没有其他解法（俺懵了）
窗口 / 联接其他表
一个公司，只要有一个人过生日都放假，其余时间上班，假设每个人每天工作量相同，问应当招多少员工可以使得一年工作量最大。
假设检验 流程描述
样本量n如何确定（需要啥数据）
1.为微信设计日报 说完 那顺序应该如何排布
2.微信营收下降了 如何分析原因

高低活跃用户/潜水用户 划分的标准？ 然后说划分不合理
随机森林是怎么样的
次日留存率计算
假设检验
检验统计量知道啥啊
p值0.7 拒绝原假设/接受原假设
特征工程 反馈 满意度极高或极低 怎么处理？

地区 品类 销售量
全国十个地区 求每个地区前三的品类
某两个地区 分别的销售量
full join 和 left join的区别
一百万数据和十万数据 如何join 谁前谁后
PYTHON
元组 列表 字典区别
列表如何增加元素
一个表 userid和买的东西 如何看什么物品经常一起被购买

毛利润 = 毛收入/GMA 降低了10% 怎么分析不同品类的影响因子
根据历史数据（生产、点赞等），如何根据留存率，确定冷启动--新手引导的内容
辛普森悖论 举例子
快手和抖音的三大区别

id 打卡时间
1.求每个id最后一次打卡时间
2.求每个id倒数第二次打卡时间
id 地址 （每个id对应不同的地址
1. 将每个id不同的地址 写在一起 逗号隔开（不会做啊）
对SQL的认识。
python
筛选函数 loc和iloc区别
平时都用python做些什么事情
选出优秀的三位同学，如何分析
金额 单数 用户活跃度
我说的归一化+权重 问 如何归一化

1、AB TEST 中用到的统计学的相关的方法你会吗，基于独立样本t 检验之类，说了我们之前在流量分层的、很多地方没有做的特别好吗，并且对于结果没有很好的定性的严谨的统计学评价
2、AA TEST 会吗，lz 直接 有点懵。这是个啥呀，现在想想 也是可以瞎猜出来的，实验变量就那么几个，如果不是AB ， AA的 话，那可以再控制的就是时间，换流量，好吧，好像查资料后我的理解并不对，附上解释和链接：
“ 在大多数其他情况下，A / A测试是一种再次检查A / B测试软件的有效性和准确性的方法。您应该查看该软件是否报告控件和变量之间存在统计上的显着差异（统计上的显着性> 95％）。
如果该软件报告存在统计上的显着差异，则说明存在问题，您需要检查该软件是否已在您的网站或移动应用程序上正确实现。”
https://www.optimizely.com/optimization-glossary/aa-testing/
3、假设检验的两类错误
HIVE:
用的多吗
常用的函数，问了order by sort by 的区别

数据倾斜原因大概会是
数据倾斜的解决方案
如果场景出现在是join 的时候数据倾斜怎么解决

介绍一个最近的数据分析项目，项目中用到了哪些指标？有没有什么结论？为什么得出这个结论
一个刚上线的短视频app，应该关注哪三个指标？为什么？
在短视频信息流app中放广告位，如何确认能否带来收益？
怎么做A/Btest?
用哪些指标判断加广告的效果？如何判断？
广告带来了每天100万的收益，但是用户的次日留存下降了，如何确定要不要上这个广告？（广告收益/用户）
sql
计算5个班级每个班级语文成绩排名前5名
计算次日留存

在空间上线性可分的两类点，分别向SVM分类的超平面上做投影，这些点仍然是线性可分的吗？给出证明（反证法 假设线性可分 …… 可得出与最大间隔矛盾）
SQL行转成列（case when）
相关性代表因果吗？为什么
Python编程（具体我忘了）
时间序列分析 ARMA模型 如何定阶 （记得要先保证序列平稳偶）
然后问了我对ARMA了解到什么程度 （我回答最后都可以写成一个特征方程，判断根来研究ARMA）
随机森林以及Lightgbm 变量重要性是如何计算的（他纠正我说这俩是一样的，其实是他错把随机森林的变量重要性与lightgbm计算一样我挺无语的）
这个组主要做因果分析，问我感不感兴趣，以及给我介绍了一下因果分析，以及他们的研究。

如何将问题转化，你认为自己研究的贡献是什么，研究取得了什么可以看得见的效果，
研究方向与导师不一样你为什么选这个导师?
问介意学一些与数分无关的语言吗，比如java，go，
如果衡量变量之间的因果关系（不需要考虑因果推断），面试过程大概 1 小时。

之前经历中最有成就感的事情，取得的结果以及中间的过程
对未来的职业规划
我的提问：PCG数分风评不好如何看待
一面面试官出现了一个技术面的答案错误，我想纠正一下
目前组里所作的业务
主要是两个人在聊天 给了我一些找工作的帮助

挖简历不再赘述，每个项目如何做的，取得的成果建议十分熟悉
用过哪些算法，对不同算法的理解
abtest 假设有两组样本各有50000个 ，单个样本服从伯努利分布 检验H0:p1 = p2
检验统计量，p值如何计算
当样本量增加时，p值如何改变
p1与p2的差值变大，p值如何改变
…… （本质把统计量的公式写出来 推一下就好了）
为了这个岗位做过哪些学习工作
我想要评价学校老师的授课情况，如何评价（量化它）
在我给出一种量化方式后，如何知道好还是不好(问题转化 如何对比)
我的提问

自我介绍
说一下关于业务的吧
如何预测淘宝双十一的销售额
淘宝的支付单数下降了，如何分析
之前做的一个项目用了深度学习的方法，如何用的，做了什么改进

自我介绍，深挖了简历中一个项目的模型，lightgbm的原理，
Lightgbm比xgboost好在哪
以及随机森林 与GBDT、xgboost、lightgbm的区别，
还有就是决策树的的划分准则有哪些（信息增益，信息增益率，基尼系数）
模型过拟合怎么办
Pandas的几个函数
写SQL

自我介绍，深挖之前实习经历，因为之前在某快消做数分，问复购的归因是如何做的？（给我一个新的场景 去下钻影响新老客购买率因素）
欠品预测建模用到了什么自变量，lightgbm模型效果如何?
提出场景问题，如果日活跃用户量下降了，如何从数据找到原因？（
1数据的准确性以及是否存在异常
2下降的维度是随时间下降，还是跟别人的差距变大了
3从几个常见维度拆分
按照新老用户的拆分；
登入平台的的拆分，比如：IOS、安卓；
按照APP版本进行拆分；
按照登录渠道的拆分，比如APP、小程序；
按照区域的拆分，比如：国家、省份；）
两道 SQL题

自我介绍，我简历中有写用图模型做商品关联分析，面试官问如何应用的？然后就是各种深挖简历，问模型效果如何，用的什么模型评价标准。
记不清了，好像又问了一些业务问题，然后问了下之后想要的职业发展方向，以及给我了一些建议。

首先是自我介绍，与数据分析相关的项目或实习经历
我讲了深度学习网络CNN算法和XgBoost算法，继续提问CNN算法的优点，
XgBoost与随机森林的区别
统计学方面：q值的含义，一类错误和二类错误
python数据分析、数据可视化、数据建模常用的包
业务方面：在流视频中插入广告，由之前的四个插一段变为两个插一段，为了评估用户反应，有什么方案进行实验
答了A/B test，用户留存率和用户刷视频时长作为考量指标
提问：收益上升，留存率不变，刷视频时长下降，怎么考虑方案选择
答了考虑收益和刷视频时长的相关关系（什么鬼qwq）
继续提问：新方案更好的情况下，老板把用户留存率作为第一考量，应该怎么应对 —— 出于业务考虑，当然是听老板的话咯，还能怎么办（应该还有什么更好的办法，但我没想出来）
最后问了怎么评估长期指标，没答出来
他给的答案是用小流量作为实验对象

1.谈一个你学到的最多的项目

2.为什么你的数据可以对标签进行预测呢？（这个问题就很。。不知道咋回答）
3.模型的X和y分别是什么
4.X中是否包含时间信息，若不包含，那是否会预测结果有偏差呢
我说对的，应该考虑时序模型
5.那时序模型比如GRU它是怎么考虑这个时间信息的呢
6.数据预处理的过程，如何进行缺失值的处理
7.刚刚提到knn插值填补数据，那如果选取的k近邻个数据也存在缺失值怎么办呢
8.随机森林的是怎么工作的，为什么要进行随机抽样
9.随机森林提取的m个特征维度是如何进行挑选的，要选哪些特征
10.如果要向用户展示precision和recall，你会选择哪个

1.自我介绍
2.做过这么多项目，说一个印象最深的
3.做项目时用到了哪些模型
4.有哪些回归模型
5.分类模型用了什么，挑一个最熟悉的说一下原理
6.PCA的原理（因为我前面提到了特征提取）
7.对聚类模型 除了K-means还知道其他的吗
8.也做过开发类项目，为什么想要去数据分析
9.数据预处理时，利用散点图等观察到了离群点之后怎么删除的呢

---

1. 时间序列模型的分析比较（AR, MA, ARCH, ARIMA, COPULA, ARCH, GARCH）分别是什么，相同点，不同点，优点缺点。
2. 文本情感分析相关的一些问题，NLP/词典，为什么，做了哪些改进，最后结果怎么样，正确率，什么指标，为什么怎么选择，目前模型有什么缺点等等
3. DNN/ CNN/ RNN/ LSTM/ RF/ DT等等分别介绍是什么，有什么联系，优缺点，异同点，什么时候选择哪种模型，分别有什么创新点，画出结构图，写出神经网络计算公式
4. 因为提到了激活函数，让我比较了一下对几个激活函数的理解，还有函数的数学形式
5. 贝叶斯理论，贝叶斯网络
6. 一些优化算法，加密算法的问题
7. 还有一些统计方面的题目，很简单
8. 笔试，10分钟两个题，我只做完了第一个，第二个口述了一下想法。都要求复杂度为O(N)，第一个是回文串，第二个是distinct longest path，挺简单的，leetcode上都有
（当时面完就觉得凉凉。。。。没想到隔了一天给了二面）

体验非常好，依然是技术面，但是题目相比之前来说实在是太可爱了，基本都没逃出上面的范围，多了一些统计概念的问题，包括样本选择，模型选择，统计检验方法选择等等。
还有就是业务题，具体不太记得清了，不过应该是和次日留存率相关的一些分析，大家分享的其他面经里有相关的问题。
还有几题SQL相关的问题。

面试官讲了一下业务范围，发展路线，闲聊了一下生活，什么时候毕业，毕业以后的打算，然后反问。15分钟很简短，但是感觉应该有戏。

介绍项目（nlp）
说一种你了解的聚类算法
俩sql题：
第一个sql题输出一个领导所有的下级

另一个sql题是窗口函数
问dau为什么突然下降然后怎么分析
hive基本知识

项目深挖挖挖（nlp+数据处理）
挂掉的笔试题重新做一遍（python）
笔试题在这里：
https://www.nowcoder.com/discuss/601720?source_id=profile_create_nctrack&channel=-1
怎么用python或者sql进行抽样
怎么用python和sql不放回抽样
了解怎么用python做数据结构吗
hive分区的用途和好处
实习时间

---

1. 自我介绍（但不知道为什么小姐姐没有深挖简历项目）；
2. 问会不会sql，简单语法什么的；
3. 让我讲讲left join，right join，inner join的区别；
4. 问了我一个业务题，但我有点没听懂 ，大体意思是我有100万的钱或者流量，然后举行个比赛，怎么去分配这个100万去得到最大的利益；
5. 问我平时看不看快手和抖音，我很诚实的说不太看，结果小姐姐说她也是；