面试记录

腾讯-看点

一轮

1. 自我介绍
2. 工作和学习中遇到的一件困难的事，如何解决，得到了什么？
3. 简历问题
   滴滴的工作内容，详细问了沉默召回项目
   k-means 的步骤，如何确定 k 值，如何确定新的族心
   介绍 ti-idf 算法
   文本分析将文本变成了多少维的向量，如何考虑运行效率的问题
4. 算法题：
   Int 64 如何计算曼哈顿距离
5. Linux 命令：如何查看文件类型
6. 数据库问题：
   数据库有哪几种常见的索引
   有用到那些数据结构，介绍一下

二轮

1. 自我介绍
2. 简历面：
   滴滴的工作内容
3. Linux 命令：工作中用过的命令
   切换用户
   Vim 文本替换
1. 常用的 HIVE 优化方法
   自己说了两种：一是计算方法的简化（修复二价）
   二是 case when 减少联表
   针对第一点扩展问了一些底层数据库的问题
   公司用了哪些数据库
   底层的储存是什么
2. 算法题
   Nums 为一数组，target 为目标值，返回数组中两相加为目标值的数下标，不能重复利用数，且找到一组就行。
   我说用递归循环
   问算法复杂度，如何解决效率问题
   用 in 效率一样
   将循环递归的思路写出来
3. 概率题
   有一原始部落，如果生的是女孩就继续生，生的是男孩就停止，问该部落的男女比例。
   1：1

腾讯-内容平台

初试：

1. 自我介绍
2. 详细问了 ABtest 项目：
   整个过程：为什么要做？怎么做？最终输出结果，给出的决策支持。
   什么是假设检验（步骤），p 值是什么，怎么利用 p 值做出判断？
   为什么使用 t 检验？
   【对假设检验问的比较细】
   有没有学过统计相关的课程？
3. 介绍经济高质量增长率这个项目
   a) 首先通过文献确定指标体系（将未知指标转为已知数据的指标）
   b) 通过各类数据平台找数据
   c) 对数据进行预处理
   d) 利用指数法计算经济高质量增长率
   e) 回归模型做预测
   什么是 prophet 模型
   缺失值处理为什么使用不同的方法：缺失值较少时使用插值法，缺失值与其他变量有比较强的关联性时使用的回归，无明显相关性时且为月度数据时使用 prophet
   ev、mae、mse、r2 各指标的含义是什么？计算公式？ev 与 r2 的差别是什么？
   svr 算法是什么？（我说找一个超平面，以及间隔ϵ，落入间隔里的点不记为损失，以外的记为损失。）
   什么是损失？（最优化的目标是使得损失最小，真实值与预测值之间的差距，不同的算法有不同的衡量方式。）
   svr 与 svm 的区别是什么？（前者做预测，后者做分类）
   LR 算法介绍一下
4. tf-idf 算法的原理（词频*逆文档率）
5. 最近学了些什么相关课程，（我说了机器学习相关）能不能挑一个算法来介绍？
   说了 k-means，介绍算法的步骤。
   k-means 的优化目标是什么？
   （我说聚类算法的优化目标是簇类距离最小化，簇间距离最大化，但是 k-means 并不是直接利用 min 簇类距离/簇间距离，它的求解是局部最优。）
   局部最优是什么？有局部就有全局，这两者的差别是什么？
   （举了函数的极值，最值的例子，在定义域里极值不一定是最值，只是局部的最值，极值就像是局部最优解，最值为全局最优解。）
6. 来到团队想要做什么？
7. 概率题（面试官自己设计的，有坑，但写完了她也没告诉我坑在哪里…）

数量	0	1	2	3	>=4

P	0.1	0.2	0.3	0.4	0

调查家庭有多少各孩子【腾讯为啥都喜欢跟孩子过不去惹】，调查了若干户结果如上，问

a) 任意一个家庭有孩子的概率（记不太清了，应该是一个概率累加的问题。）
b) 随机取出一个孩子，他有姐姐或者妹妹的概率。（有坑的题，面试官带我算了一下）
   0.3*0.5（0.5 是表示女孩的概率）+0.4*（1-0.5*0.5）
   面试官说孩子来自家庭中的哪个是不确定的，上面的方法是不对的，然后结束了。
   后来想了一下：0.3*0.5*$C_2^1 0.5^1 * 0.5^1$+0.4*（1-0.5*0.5）**$C_3^1(\frac{1}{3})^1 * (\frac{2}{3})^2$，但不知道是不是这么个意思？

8. Sql 题，调查了 n 个家庭孩子的数量情况，写个 sql 变成上述概率题展示的样子。
   Id 表示家庭标号，kids 表示孩子数量。……
9. 最后问了我还有什么需要提升的地方
   大概就是对简历的熟悉度。
   招人的标准，要么某一方面特别强，要么学习能力出众。
10. 整体感觉：
    面试官问的特别的细，对简历挖的蛮深的，比较看统计学知识、概率论知识以及机器学习方面，所以简历上写的东西深至原理、公式，都要懂！

复试：

1. 自我介绍
2. 问了爱奇艺、滴滴的工作：
   - 广告的流程介绍一下
   - 收入异动是怎么分析的？
   - 怎么拆解指标的，指标异动的标准是怎么确定的？有没有置信度？
   - AB 实验的工作内容

   主要是针对实习的一些细节问了，关注实习工作中各类指标或做事的科学性。面试官很 nice，更像是聊天探讨的感觉。
3. 如果我加入，我的工作内容？
   - 指标体系的构建
   - 归因分析（指标异动）
   - 推断分析（AB 实验）
4. 总体感觉
   二面更多是对实习工作内容的考察，看你是否真的参与过这些工作吧，更偏向业务一点。
   时间不是很长，大概 40min，聊天为主，考察为辅的感觉。

HR 面：

1. 基本信息的确认
   学历，是否是保研
2. 硕士研究方向，具体做什么
3. 在爱奇艺学到了什么？
   拆解问题的思维
   系统做事的思维

4. 为什么 19 年去做了产品经理？
5. 是否有其他 offer？
6. 是否有亲人在腾讯？

阿里-蚂蚁集团

1. 自我介绍
2. 看你的专业比较偏系统，为什么想做数据分析？
3. 介绍一下自己参与过的一个项目（why?how?result?desicion?）
   爱奇艺的 abtest
   然后对这个项目深挖了
   有没有继续深入分析，广告收入总体减少的原因？
   因为人均观看时长增加了，但广告收入减少，是不是广告渗透率原因巴啦啦？
   我说是一方面时点击率低，优化广告素材，另一方面是竖版广告的单价低，拉动大广告主投竖屏广告。
   总之，面试官觉得之后的广告收入降低的原因深挖不够，没能说服他。
4. 说一下收入监控怎么做的？
   横向定位模块，纵向定位原因，巴啦啦说了一大堆。
   面试官觉得这些可以通过看板，bi 自动实现，那么这其中数据分析师的作用是什么？
   一方面广告业务的逻辑较为复杂，看板没能把所有逻辑一一展现，需要从底层表取数分析，另一方面会有看板之外的突发原因，例如风控因素带来的广告主投放减少等。
5. 分析支付宝里理财界面 DAU 下降的原因
   分内部原因与外部原因
   内部原因：分析总体流量与各分页流量情况，如果是普降，考虑是否在正常范围以内（可以通过同比判断），如果不在正常范围以内，考虑是否有技术问题或其他问题（活动啥的）；如果只是理财页面流量下降，分析理财页面的是否存在技术问题或其他。
   外部原因：节假日时间因素，例如过年期间大家对理财关注更多些，流量又增加，现在减少了；市场行情因素，行情较差，部分人从理财市场撤出来，流量减少。
   【感觉自己对 DAU 这种指标的分析真的是不拿手…胡编乱造，说的不是很有逻辑。】
6. 简单介绍一下滴滴的实习经历
   简单说了一下，没深挖，转而问起头部互联网的商业逻辑。
7. 说一下你理解的头部互联网的商业逻辑
   当时就懵了，反正就随便说了些，抖音跟拼多多（社群裂变、用户下沉，高年龄阶段人群，消费心理，微信的流量入口）
   当我说出拼多多我就后悔了，我就知道面试官肯定问拼多多与淘宝之间的啥啥啥的，什么叫自己往火坑里跳。
8. 拼多多与淘宝的优劣势。
   …大脑一片空白以至于忘了自己说了些啥。
   讲了淘宝的双十一，建立一个品牌，还有双十一晚会，将一个日子成为一个购物的象征。
9. 最后问了需要提升的地方。
   面试官说了我的优势是有两段相关的互联网实习经历。
   缺点是缺少商业思维…缺少对项目的深入了解
   然后给我科普了一下目前互联网的数据分析干些啥：

a) 偏大方向的决策：服务于部门的业务目标，指定业务的 kpi 啊（要赚多少钱，要有多少流量等等），并分析指标是否合理。
b) 偏日常运营的决策：指标监控，各类产品、运营活动是否完成了业务目标，用数据佐证。

10. 总体感觉：
    第一次遇到这么业务的数据分析面试，对商业逻辑这方面确实不是很懂，其实对头部互联网的理解也比较少，之前总觉得产品才需要了解各互联网公司的产品矩阵、商业化方式等等，现在看来偏业务的数分同样需要，这方面的知识需要补充一下。
    面试时间很短，37min，感觉只是一个简单的筛选面试。

阿里-淘宝

1. 自我介绍
2. 工作中觉得最值得拿来说的一件事？
   AB 实验
   这里讲了很久，面试官甚至打开了爱奇艺 app…很细节地问
3. 为什么觉得是这件事最值得拿来说的？
   独立解决问题
   理论应用于实践
   工作思维的培养
4. 目前人生中觉得自己遇到的最大挫折。
   …
5. 有没有用过机器学习方面的东西？做这个的流程，用了什么模型，怎么评估模型，怎么选择模型等等？
   当时说了经济高质量增长那个，只说了最后模型预测那部分，感觉挺干巴的，以后感觉可以说文本分类那个，好好准备一下。
6. 反问：这个岗位的业务，具体做什么的？
7. 总体感觉：没面多久，大概 30min，感觉不怎么样，也没能从面试中了解到面试官想考察什么。

字节-广告

1. 自我介绍
2. 如何进行收入监控的？
   巴拉拉说了一堆，横向+纵向，流量侧+供给侧

   面试官反问，如果不止一个横向位置发现了问题怎么办？
   去掉长尾效应的地方，占据了大多数变动的所有的位置都需要进行分析说明
3. 怎么判断是波动还是异动？
   业务标准
4. AB 实验的项目：

巴拉拉说了一堆，也被反问了一堆问题。
有一个没回答上来：如何保证两个假设检验的弃真错误？显著性水平应该是 95%*95%，而不是 95%。
emmmm，说实话，理解不是很深刻，但感觉好像是这么意思。

5. 关于广告行业的一些基础知识
   ecpm 怎么计算？
   竞价逻辑是什么？
   广义第二高价，面试官问我，为什么选择第二广义高价这种逻辑，选择最高价不是更适合赚钱吗？
   之前学业务的时候了解过，但给忘了…
   博弈问题，要看看…
6. 滴滴的工作，简单介绍一下沉默召回项目。
7. 如何建立的指标体系？
   原有的指标体系加上项目特有的考察指标。
8. 解释什么是沉默召回率。
9. 针对沉默召回项目，你觉得还应该考虑什么指标？
   激活时间，再次沉默时间、成本、收益等
10. 为什么选择用 rf：
    数据量大，rf 在分类问题上表现较好
11. 你做特征输出的时候使用的是 rf，其他人在做预测时用的是什么模型，模型不同不会导致出现偏差吗？
    时间太久了，其实这些我已经完全不记得了，编不好，感觉以后还是从简历里删掉。经不起深挖。
12. 随机森林的损失函数？
    当时一直在想损失函数。
    而后面试完了，想想，随机森林有损失函数吗？不是信息熵的原理？
13. sql 题
    a 表：id, key1
    b 表：id, key2
    找出 id 只出现在 a 表的 id, key1
    用了 not in，面试官说效率低
    用了 where a.id != b.id 面试官说 hive 不支持
    然后用 left join, 在 where b.id is null， 其实很简单，但是当时脑子短路，花了几分钟才写出来…
14. 概率题：

一枚硬币抛 100 次，如果正面朝上 60 次，能否认为这枚硬币正面朝上和反面朝上的概率，正面朝上更大（或者不同）？

说了一堆，但是没能理解到这题的意思。
面试官提示是假设检验的问题。
总之，中间很曲折，让我明白了是关于总体比例的假设检验问题。
然后我说了正态分布近似二项分布，计算那个统计量，当时自己也没怎么说明白，p 来 p 去，搞不清楚是哪个 p 了，面试官又说不用近似，所以有点懵。

$$\frac{\bar{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

然后面试官又说到 p-value 了，问 p-value 的含义：
emmm，有一种你知道它是个啥，但是却无法具体说出来的感觉…
H0 是真，但错误拒绝的概率？

面试官想要考察的应该是知乎的这个问题， P(60<=X<=100) > 0.05 ,不拒绝原假设。
面试官一直强调的 P 值也就是 P(60<=X<=100)这个!
https://www.zhihu.com/question/23149768

15. 总体感觉：
对统计学考得比较深，而自己对假设检验这边理解不够深入，还是停于表面了。
还需要深入了解一下假设检验！

美团-骑行事业部

一轮
1. 自我介绍
2. 爱奇艺监控收入，如何进行，如何设计指标进行自动监控？
3. 滴滴实习沉默召回项目：什么是沉默召回率？如何衡量项目效果？
4. sql 题：

order 表
字段：userid, orderid, bikeid, starttime, endtime(日期格式为 yyyy-MM-dd HH:mm)
分区：dt
bikes 表
bikeid

a) 求 2020 年各个月份的订单数、骑行用户数、被骑车数

```
select
month(dt) as mon
count(distinct orderid) as order_num
count(distinct userid) as user_num
count(distinct bikeid) as bike_num
from order
where dt between '2020-01-01' and '2020-12-31'
group by month(dt)
```

b) 周转率=订单数/车数，求 2021 年 4 月 7 日的周转率分布

```
with tab as
(
    select
b.bikeid as bikeid,
count(distinct orderid) as order_num
```

```
        from bikes b
        left join order o
        on b.bikeid = o.bikeid
        where o.dt = '2021-04-07'
        group by b.bikeid
)

select
order_num as `周转率`,
count(bikeid) as `车数量`
from tab
group by order_num
```

c) 4 月 7 日空闲时间 空闲时间的数量

```
select
free_times,
count(order_id) as free_nums
from
(
select
order_id,
(lead(endtime, 1, null) over(partition by bikeid order by starttime) – endtime)/60 as free_times
from order
where dt = '2020-04-07'
) a
where free_times is not null
group by free_times
```

有 null 参与的加减乘除如果没有处理过，结果都是 null

二轮

业务面，总体感觉比较简单。

1. 自我介绍
2. 学校中学习的理论知识与实际应用的有没有感觉到什么区别？
   - 理论知识是顺着下来的，例如介绍知识点，然后通过数据，展现知识的具体应用，而在实际中往往是逆向的，出现一个问题，通过对问题的拆解，去寻找合适的指标来衡量问题，然后通过指标找到需要的数据。
   - 数据量的区别，学校接触的项目数据量一般较少，而工作中的数据量通常很大。
3. 通过一个工作中的一个具体例子来说明以上的问题。
   爱奇艺 ABtest，数据选择的问题
   怎么发现问题的？——之前有预期，符合常识情况的预取。
4. 介绍一个学校的项目：

经济高质量增长（简单的步骤）
你的研究方向是什么？为什么跟经济有关？
做预测的模型是如何进行选择的（选择适合小样本的预测模型）

5. Python 填写的是一般？
   因为虽然系统学习过语法，但是没有经常使用。
6. 平时的爱好，喜欢的体育运动。

快手-流媒体

一轮

1. 自我介绍
2. 做 sql 题：
3. 假设检验的步骤
4. P 值的含义
5. 假设检验的分布如何选择？
6. 实验组与对照组，但是实验组命中策略的人只有一部分，那么怎么进行假设检验？
   这题没回答出来，面试官提示，可以使用实验组命中策略的人群以及在对照组中找到相似人群（例如通过人群扩散等方式），然后进行假设检验。
   人群扩散这个概念第一次接触，查一下。
7. 介绍工作中的一个项目
   爱奇艺的 ABtest（详细）
8. 对机器学习的熟悉程度？
   知道基本原理及调包实现。
9. 什么是过拟合？什么情况会导致过拟合？如何解决过拟合？
   造成过拟合的原因：
   训练程度过高
   训练样本过少
   模型过于复杂
   样本特征过多

   如何解决：
   特征工程
   降低模型复杂度
   给损失函数加上惩罚因子，正则化

10. 分类中正例与反例样本不平衡，如何解决？
    欠采样
    过采样
    阈值移动
11. 卡顿率增加？如何分析？
    内部原因：appbug？有不知道的新策略影响？
    外部原因：设备（手机型号）、APP 版本、系统、地域、时间段、节假日、网络（wifi…）、运营商等等

龙湖集团

一面
自我介绍
Xgboost 的基本原理以及与 gbdt 的区别
Spark sql 排序函数
生活中遇到的最大困难
理想的工作状态

京东-零售

一面
自我介绍
爱奇艺收入分析项目
学校经济高质量增长项目
其他纯聊天

二面
自我介绍
其他纯聊天，考察去京东的意向

字节-互娱研发

一面
1. 自我介绍
2. 爱奇艺收入分析的工作内容。
3. 爱奇艺 abtest
   基于此，问了 t 分布与 z 分布的使用场景。
   点击率与留存率的检验应该用什么检验？
   面试官说点击率是不独立的，而留存率是独立的，所以对点击需要进行类似于过采样的方式？
   例如说 banner 的每个轮播图的点击率之间是由相关性的，如何对这些点击率进行假设检验？
   【最后也忘记问，应该具体怎么做？如何进行过采样？】
4. 实验设计：
   开 AABB 好还是 AB 好？
   开了 AABB，结果显示 AA 没区别？但是 B1 显著正向，B2 无明显区别？为什么？如何分析？

我回答：可能是没有真正做到人群同质，选取人群特征进行分析，找出 4 组人群中明显差异的人群特征。

面试官继续：如果去除上述的原因，继续开实验，你会怎么开？

我说了 ABB，【现在想来应该开 AB 就够了？已经在一定程度上保证了人群同质？】

面试官问：新开的实验需不需要调整显著性水平？

【没有回答对，后面问面试官时，面试官提示是多组实验，进行的是多个两两之间的比较，犯错误的概率时大于只开两组实验的比较。跟之前暑期面字节时提出对爱奇艺 abtest 显著性水平质疑一样。】

例如说开 AA，两个不一样的概率为 5%，而开 AAA，任意两组不一样的概率为$C_3^2 * 5\%$

5. 机器学习相关:

   Xgboost 原理

   偏差、误差的区别，如何判断一个模型的好坏？【偏差与方差】

   如何解决样本不平衡

   随机森林与 xgboost 哪个泛化能力更好？

6. Hql：

   有没有接触数据倾斜？

   Mapreduce 原理 map 干吗，reduce 干吗

   窗口函数接触过哪些？

   还有个问题：好像是什么 shap value，没听清……

7. 概率题：

   两个均匀分布

   - [0,2] P(x<= 1)=? P(x=1)=?
   - 学校【矩形】周围停车不能超过 2 小时，管理员绕学校一周 2 小时（匀速），管理员看到一辆车，会上去打上标记，绕学校一圈后，如果车辆还在，就贴罚单。

     问：现在需要停 3 小时，被贴罚单的概率？（1/2）

     4 小时？（1）

二面

1. 自我介绍
2. 介绍一个分析相关的项目：

   爱奇艺收入异动拆解

   问了跟滴滴一样的问题：如何分析是哪个因素影响更大？影响占比？

   我说了 shap value 的思路，面试官表示如何求解权重？因为收入是乘法，而 shap value 的思想是加法。

   然后说了敏感性分析，控制其他因素不变，其他一个因素改变，对总体指标的变化。面试官说这个思路可以。【需重新看看敏感性分析的内容】
3. 介绍一个建模项目：

   说了原创权益挖掘

   针对这个项目问了一些细节：

   如何进行特征选择？

   Auc 的原理？

   项目里为什么追求准确率？

   Catboost 原理【没回答上来…】
4. 实验设计，考点是正交流量，详细看一下司南计划！

   两个实验要在同一组流量上开？怎么开？

5. 假设检验相关：
   非参数与参数检验
   A 组男女的付费率高于 B 组，但是整体付费率低于 B 组，为什么？辛普森悖论【记一下详细定义以及数据样例吧】
6. 写 Sql，取不同用户对不同主播最近的一次打赏时间与金额【排序函数】
7. Python 编程：因式分解，逻辑很简单，但是写得不熟练
8. 问题：机器学习的原理；python 不够熟练

三面

1. 自我介绍
2. 介绍一个项目，要说出这个项目的价值、分析思路、以及在这个过程中遇到的困难。
   说了腾讯原创挖掘项目
   四个部分：背景、数据准备、算法流程、工程流程
   细问了一些：具体带来了多少价值。
3. ABtest 相关：
   Abtest 的原理
   P 值是什么？
4. Sql 题
5. 指标设计，如果给 b 站设计 5 个核心指标，选什么？为什么这么选？
   产品定位：B 站 ugc，平台型产品
   消费指标：DAU，日人均消费时长
   供给指标：日均投稿数
   性能指标：bug 率
   盈利指标：日人均消费金额
6. 在美国投放了类似 b 站的产品一年后，我们向墨西哥投放，投放两周后我们发现，产品安卓端的 7 日留存比 IOS 低 15%，而美国只低了 3%，分析其中的原因。
   - 详细数据，若 ios 端用户数据量本身就小，导致其留存较高【比较两国在各端数据的详细情况】
   - 在各端细分人群分析，挖掘是否是人群差异导致
   - 考虑是否存在新奇效应的影响【数据还没稳定】
   - 细分安卓端手机厂商留存差异，进一步分析差异导致的原因
   - 细分安卓端渠道差异，不同渠道用户群差异？
7. 职业规划

滴滴-顺风车

1. 自我介绍
2. 说一个腾讯的印象深刻的项目
   原创 cp 权益挖掘【感觉逻辑还要梳理一下，没有很流畅】
   针对这个问了：
   特征工程
   类别不平衡如何处理
   有没有尝试过其他模型，效果如何。

3. 写 sql 题
   Max() + group by
   排序函数取每组前三 ： 1，1，3：1，2，2（4）：1，1，1，（4）【用哪个排序函数需要考虑一下】
   有用户 id 以及其登录时间，取每个用户最大的连续登录天数【没想出来当时】
   sql 题给的思考时间都挺短的，只要求思路。
4. python 编程
   一行代码实现，0-100 中所有能被 7 整除的数以及包含 7 的数
   匿名函数：Lambda (如何实现？)
   v = list(map(lambda x: if x%7 = 0 or (x-7)%10 = 0 or x//10 = 7, [i for i in range(0,100)]))
   尝试一下。
   列表生成式：[x for x in range(0,100) if x%7 = 0 or (x-7)%10 = 0 or x//10 = 7]
5. 爱奇艺收入拆解项目，整个流程
6. 爱奇艺 abtest
   针对这个问了
   如何确定样本数量
   如何确定哪个因素对收入波动影响最大（我说了敏感性分析），面试官提示 shap value
7. Prophet 模型原理，参数有哪些？跟传统时序模型或者线性回归相比，有什么优势？
8. Tableau 的复杂函数 fix 函数？？？ 反正没听过
9. 最后问了需要提升的地方
   Sql 不够扎实
   Python 编程基础太弱，说关于封装的都没问…
   Tableau 不够熟悉，Tableau 功能很强大…不是我理解的做做表，做做图…
   机器学习关于特征选择那块比较薄弱
   Prophet 模型写了就要能说

小米-大数据分析部

一面：【挂】

1. 介绍一个在腾讯的一个项目：
   原创权益挖掘
   从四个部分来说：背景、数据准备、算法侧、工程侧
   细节问题：**catboost 的原理**
2. 爱奇艺的工作内容
3. **tf-idf 的计算**
4. **word2vec 的原理：**
   有两种方式：CBOW 与 skip-gram，如何生成向量
5. 逻辑回归的原理
   逻辑回归减少损失的方法
   随机梯度下降与批量梯度下降
   如何解决样本的不平衡
6. Bagging 与 boosting 的区别

7. **ffm 与 mf（二维交叉特征）**
   完全不懂…
8. pms 原理
9. scala 的算子？
   不太懂，后来简单问了一些 dataframe 的操作
   分布式机器学习用的哪个包，xgboost4j
10. 深度学习：不懂，跳过
11. 写题，冒泡算法，居然没写出来…

B 站-数据挖掘

1. 计算机网络知识：DNS
2. Hdfs 原理、mapreduce 原理、hive 与传统数据库的区别
3. 类与对象
4. 大数定理
5. 行转列与列转行，sql lens 函数
6. 爱奇艺项目
   如何设计实验组：哈希分桶、空跑、AABB

贝壳-数据挖掘

1. 自我介绍
2. 算法题：最长连续子序列
3. 腾讯的工作内容（用了什么特征、auc 是什么？有没有做特征交叉）
4. 爱奇艺做 ab 实验，ab 实验原理、如何做假设检验
5. rf 原理，有放回抽样
6. 数据倾斜造成的原因以及如何解决
7. Sql 题：窗口函数
8. Did 原理
9. Tf-idf
10. 经济高质量增长项目：为什么选择不同的缺失值处理方式？

字节-互娱

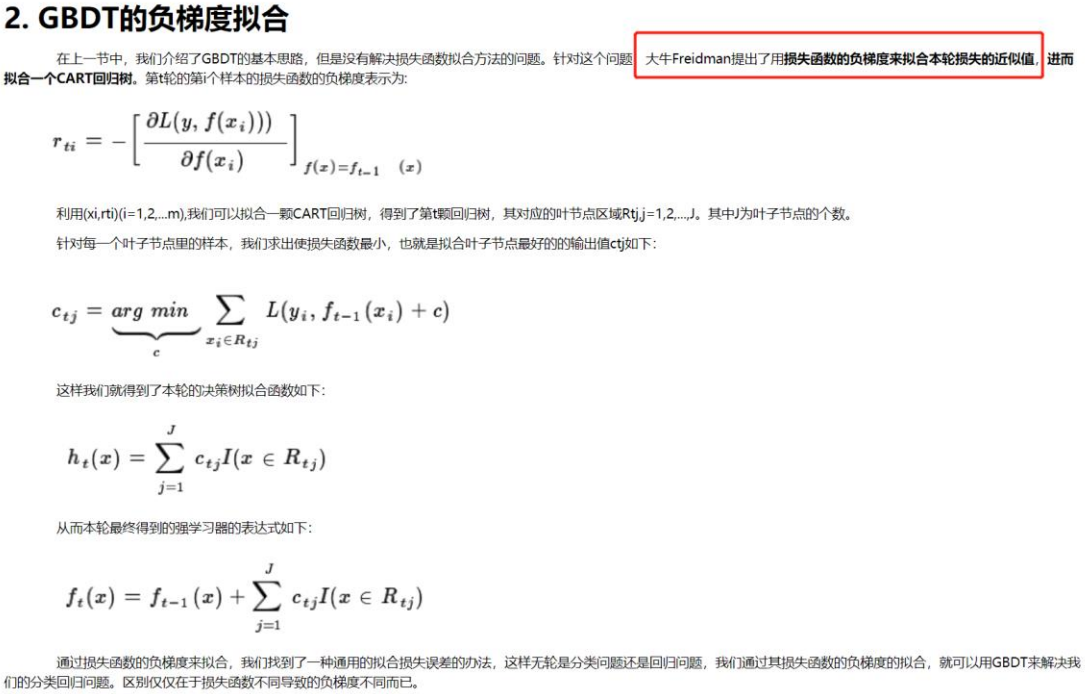
二面：
1. 自我介绍
2. 原创权益挖掘项目（问得很细，特征选了哪些特征，如何进行特征工程，阈值选取的合理性？）
   针对这个项目细问：如果我要挖掘有潜力的主播，应该选取哪些特征？

刚开始我就直接开始回答用哪些特征，但其实面试官更想听思路，如何选取特征？

- 先通过无监督的机器学习对主播进行分群，分析每个分群的主播的一些特征，找到有潜力主播的分群，分析其特征，然后利用这些特征进行有监督的机器学习建模。
- 分析已有的高质量的主播在还未成为高质量主播之前的一些特征，利用这些特征进行建模。
- ……

面试官说整个过程就很平淡，大家都是这么做的，没有体现我的思考。

3. 归因分析的案例：说了爱奇艺的收入监控异动分析。
4. Xgboost 的原理：
   为什么 xgboost 可以实现并行？
   为什么叫梯度提升树？（因为拟合残差就是负梯度）为什么残差就是负梯度？（梯度本身是指损失较小的方向，而残值是与真实值与预测值的差，用损失函数的负梯度来拟合本轮损失的近似值。）



2. GBDT的负梯度拟合

在上一节中，我们介绍了GBDT的基本思路，但是没有解决损失函数拟合方法的问题。针对这个问题，大牛Freidman提出了用**损失函数的负梯度来拟合本轮损失的近似值，进而拟合一个CART回归树。**第t轮的第i个样本的损失函数的负梯度表示为:

$$r_{ti} = -\left[\frac{\partial L(y, f(x_i)))}{\partial f(x_i)}\right]_{f(x)=f_{t-1}\quad(x)}$$

利用(xi,rti)(i=1,2,...m),我们可以拟合一颗CART回归树，得到了第t颗回归树，其对应的叶节点区域Rtj,j=1,2,...,J。其中J为叶子节点的个数。

针对每一个叶子节点里的样本，我们求出使损失函数最小，也就是拟合叶子节点最好的的输出值ctj如下:

$$c_{tj} = \underbrace{arg\ min}_{c} \sum_{x_i \in R_{tj}} L(y_i, f_{t-1}(x_i) + c)$$

这样我们就得到了本轮的决策树拟合函数如下:

$$h_t(x) = \sum_{j=1}^{J} c_{tj} I(x \in R_{tj})$$

从而本轮最终得到的强学习器的表达式如下:

$$f_t(x) = f_{t-1}(x) + \sum_{j=1}^{J} c_{tj} I(x \in R_{tj})$$

通过损失函数的负梯度来拟合，我们找到了一种通用的拟合损失误差的办法，这样无轮是分类问题还是回归问题，我们通过其损失函数的负梯度的拟合，就可以用GBDT来解决我们的分类回归问题。区别仅仅在于损失函数不同导致的负梯度不同而已。

5. Python 写题：
   split() set() sort()

三面：

1. 自我介绍
2. 腾讯原创权益项目：问得非常详细
   有没有做数据大盘的摸底？你们捞的池子里原创比例有多少？了解项目的天花板。
   为什么预测的流量入口不是全量，然后卡阈值比较高的送审？
   Badcase 有哪些？怎么根据 badcase 做模型的调整？
   特征选了哪些特征，问得比较具体。
   过滤特征怎么做的？
   训练数据由哪几部分组成？如何分配权重？怎么保证样本分布与大盘相似？

Xgboost 的原理。
Xgboost 如何处理缺失值。
Xgboost 的正则方式：L2 + 叶子结点数量
Xgboost 结点分裂公式，以及 gamma
Xgboost 特征重要性的计算方式？其他特征重要性计算的方式有哪些？

1. 两组长度不同的时序序列，如何计算其相似性？
2. 一个数组 有两个操作：任意位置加一 以及整体加倍，然后给定最终状态，求全是 0 的数组到最终状态需要的最小步数

美团-到店

一面
1. 自我介绍
2. 原创首发整个项目：如何解决数据口径问题
3. 有没有接触数仓、spark、hive 相关的底层技术
4. 遇到的困难：技术层面与非技术层面
5. 数据倾斜、sql 优化
6. Spark 算子的区别，rdd 与 dataframe
7. 实习的增量学习部分
8. Sql：连续登录
   Left join and 与 left join where 的区别
9. 自己的三个优点

二面
可能我的背景并不是很吻合，问得都是偏软能力方面的：学习能力、解决问题的能力等。

Hr 面：
1. 介绍在腾讯的工作内容
2. 为什么不转正
3. 你觉得架构变动对新入职的员工有什么影响
4. 怎么理解数据开发与数据分析的 gap，准备怎么学习
5. 学习能力比较强，有没有具体的例子
6. 听你的描述你的 mentor 是引导式的指导风格，你更喜欢哪种领导风格？
7. 经常实习是因为实验室没有活？
8. 为什么说自己大三大四就确定了职业方向？
9. 实验室项目带给你的成长？

虎牙-经营分析

一面

1. 自我介绍
2. 群体可解释性，详细问
3. 爱奇艺画像分析，一般会从哪些画像纬度分析：
   基础社会属性、搜索画像、观影画像、基础设备画像等
4. 算法题：有个 0-6 的整数均匀分布的函数，现在想要 0-20 均匀分布的函数，怎么办？
5. 过拟合如何解决？
6. xgboost 与 svm 谁更容易过拟合？
7. 有两个特征，x 的平方+y 的平方 =100，标签为 1，反之为 0，而且有很多样本，问：能用 xgboost 建模吗？用能 svm 建模吗？
   【感觉是在考察模型能否解决非线性问题，svm 高斯核，映射到高维空间】
8. 如何看待置信度
9. ab 实验设计相关

二面

1. 自我介绍
2. 腾讯原创权益卡挖掘
3. xgboost 与 gbdt 的区别
4. 爱奇艺 abtest 项
5. 假设检验的步骤
6. 论文做了什么，怎么做的（第一次被问论文
7. lpl 赛事对平台的影响如何刻画
   用户量（新增、日活）
   盈利（直播打赏、引入新用户之后的现金转换）
   人均观看时长
   投稿量等

   之前的回答是从几个关注的方面去看的，现在思考感觉更像是一个完整的指标体系的搭建，首先去建立一个北极星指标，然后对北极星指标进行拆解。
   对于这种内容平台

8. 赛事直播在线人数较去年下降，分析为什么？

三面

比较偏向于聊天，没有问太多的项目以及专业知识。
了解我在腾讯的一个工作状态，以及工作环境，部门，以及一些跨部门的项目了解。
如何看待学校与实习带给你的东西？技术的更新迭代很快，学校一般很难 follow。
城市偏好。
反问：看重候选人什么品质：自驱力。

Hr 面

随便聊聊，考察意向

oppo

一面：
1. 自我介绍
2. 腾讯原创权益挖掘
3. 爱奇艺收入分析
4. 各段实习有什么区别（递进）
5. 不足之处：不够专精
6. 职业规划：互联网+数据分析

二面：

当时忘记记录了，忘记了…

网易云-数据分析

一面：【挂】
1. 自我介绍
2. 腾讯项目介绍，问了一些业务问题
3. 爱奇艺项目介绍：收入异动分析
4. 说一个自己独立做的项目：站内外高粉看板分析
5. 职业规划：偏算法还是业务
6. 对城市的偏好

京东-零售

一面：【挂】
1. 自我介绍
2. 群体可解释性项目是什么？介绍以下。
3. 群体可解释性中 IP 识别不准确 对聚类有什么影响，如何解决这样的问题？
4. 原创权益挖掘项目，模型效果如何？有没有达到业务要求？没达到为什么上线？
5. 偏算法还是业务：

滴滴-国际化

一面：
1. 自我介绍
2. 广告收入异动分析

3. 你知道广告收入的体系为什么要这么搭建吗？
   不知道…
4. 原创挖掘项目，针对项目问了些问题。
5. 给自己的 python 与 sql 打分
6. 设计一个外卖 app 指标体系
7. 最小商机密度，也就是店铺跟人群 match，怎么计算这个最小商机密度，判断某个地方是否开放外卖业务？
   店铺数量、类型
   人群偏好（用户喜好分析）、数量、频次
   骑手数量、送货能力
   三者匹配，匹配规则暂定为设定阈值等
8. 如果现在想开一座新城的外卖，你会怎么选？
   先找有没有相似城市，分析相似城市的成功经验，判断是否能进行复用。
   如果没有相似城市，可以利用之前说的最小商机密度，在城市做试点，由小到大。

反问：不足之处
Stroying thinking ，结构化思维（认为我只有讲腾讯项目的时候比较有条理）
对指标体系了解太少。

二面：

1. 自我介绍
2. 原创权益挖掘项目
3. 爱奇艺收入异动分析
4. 某公司负责某银行的 ATM 机运营，如何让公司老板了解整体的运营状态？**【指标体系搭建】**
   选指标：现金流（流入、流出）；各机器人均等待时长
   现金流的流入拆解：在线机器量、存钱频次、次均存钱量
   现金流的流出拆解：在线机器量、取钱频次、次均取钱量
   各机器人均等待时长
   分维度：分区域，区域下分机器
   为什么选择这两个指标：
   现金流了解现在的机器是是否满足用户的基础需求。
   等待时长了解衡量用户的体验。
5. 房地产中介进行二手房的销售与租赁，现该中介有一家店铺业绩不好，分析其原因。**【异动分析】**
   外部：
      政策管控
      地域、交通、是否是学区等，对房价的影响
      竞争对手
   内部：
      先分模块：交易房产额与租赁业务额
      漏斗拆解：交易额 = 付费用户 * 单价
         付费用户漏斗：可达用户量 → 沟通用户量 → 看房用户量 →付费用户量
         单价：自然原因导致单价较低还是定价策略有问题
6. 外卖业务，如何挑选商铺进行签约？对店铺进行优先级策略设定。**【策略制定】**

商铺特征：规模、品类、品牌、热度、好评度、承接业务能力…
用户：用户偏好、人群分布（距离）
骑手：骑手分布
将用户、骑手的影响因素作为权重调整在品类、距离特征上，然后可以利用聚类分析对商铺进行聚类，选取高优先级商铺。
面试官说不用很复杂，反问我选择外卖时关注的是哪些维度：口味符合、好评多（销量、好评）、距离。
策略的选择就可以根据用户需求来。

7. 反问：提升点：商业思维《营销管理》

Hr 面：

1. 自我介绍
   学历
   实习经历简单介绍
   兴趣爱好介绍
2. 会日语吗？不会…
3. 对工作的了解程度
   工作内容
   工作要求
   外卖了解
   国际化了解
4. 实习学习中，成长最大的项目？
   独立解决问题能力
   结构化思维养成
   技能学习
5. 相比于其他公司，为什么选择滴滴选择这个岗位？
6. 更喜欢哪种领导风格？
7. 英文聊天：别人怎么描述你的性格，你想改进的性格，这个性格好的与不好的点？
   要命…
8. 反问：工作状态、日常工作、英文要求