业务

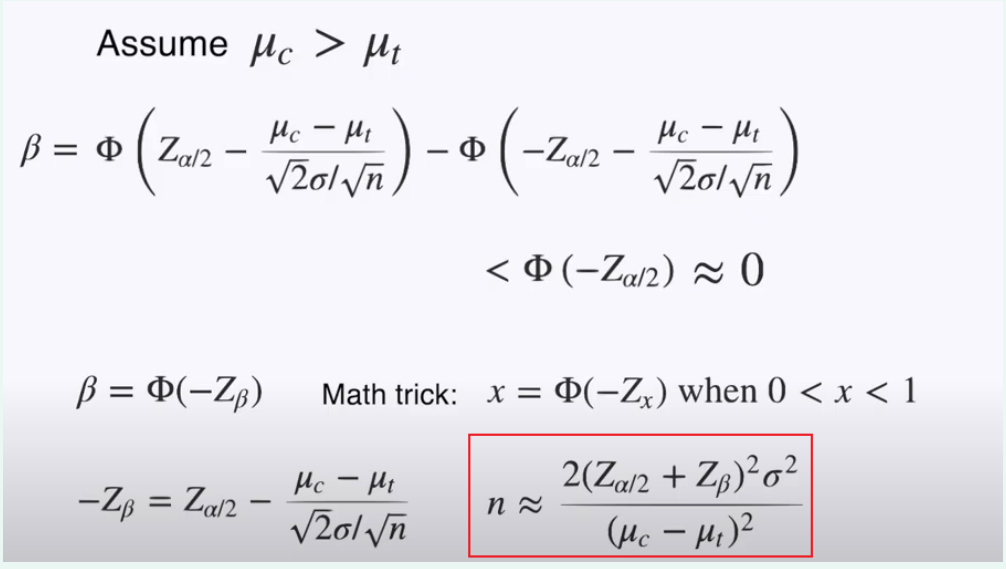
- [ ] A/B test
  - [x] A/B test怎么分流（抽样）

    > 互斥，正交（正交要解释清楚）？
    >
    > 原因：为了解决辛普森悖论的问题，不同方案所抽取的用户应该同分布，而且能够代表总体的分布，比如A有20男，那B也应该有20男。**参考**，
    >
    > 因为abtest中唯一可以变的是方案，所有其他条件都应该一样，包括用户的分布。

  - [ ] 搜索框改为搜过按钮，预测视频播放量的影响
  - [x] a/btest的使用条件

    > A/Btest之前如果有问题，如：数据损坏、数据混乱，抽样不随机。需要先做aatest，即相同的方案在不同用户上的体验，以确定用户抽样的随机性

  - [x] a/btest的样本量怎么确定，**参考1**，

    > 一般有“两个总体均值”和“两个总体比例”两种情况。
    >
    > $n_1 = n_2 = \frac{z^2_{\alpha/2}(\sigma_1^2+\sigma_2^2)}{E^2}$，$n_1 = n_2 = \frac{z^2_{\alpha/2}[\pi_1(1-\pi_1)][\pi_2(1-\pi_2)]}{E^2}$
    >
    > 在实际场景中，常有α, β, E三个参数，公式为
    >
    > $n = \frac{\sigma^2}{\Delta^2}(z^2_{\alpha/2} + z^2_{\beta})$，
    >
    > 具体推导方法：**计算样本量**，或如图所示，假设取2σ，$z_{\alpha/2} = -1./96, z_{\beta} = -0.84$，那么最后就是约为16。
    >
    > 
    >

  - [x] abtest需要做多少天？【见abtest流程】

    > https://vwo.com/tools/ab-test-duration-calculator/

  - [x] a/btest的截止条件，即什么情况下可以停止试验（除了统计学方法除外）【见abtest流程】
  - [ ] a/btest的局限

> 尽管设计简单，所需样本量小，但abtest一次只能测试一个特征。可以考虑多变量测试

- [x] 自己想做的a/b test

> 微信读书，最近读过的书放在左上角还是右下角。

- [x] 试验结果统计显著，实际不显著，为什么?

> 一个可能原因：样本量太大，和总量差别较小。比如APP启动时间优化了0.001秒，在统计上可能是显著的，但用户感知不到。没有太大的实际意义。

- [x] 试验结果统计不显著，怎么判断收益

> 一种通用的方式：将指标拆分，每天都观察。如果变化曲线每天实验组都高于对照组，即使统计不显著，也认为在这样一个观测周期内，实验组关键指标的表现是优于对照组的，即可以上线

- [x] 如果核心指标显著提升，一定能上线吗?

> 如：提高帧率，但会增加耗电。一个方面的优化可能会导致另一个方面的劣化。性能提升时也可能对其他部门造成影响，从而影响收入。

- [x] 开ab测试是需要成本的，你不觉得每一次都开ab试验，成本会过高吗?

> 如果只是验证一个小按钮或者小改动，可以在界面上设置一个开关，用户可以自行决定采用哪一种方式。最后可以通过开关的相关指标来判断用户倾向于哪一种形式。也可以做一些问卷之类的。

- [ ] abtest如何抽样保证分布相似(我不会,我弱弱的说了句哈希?面试官立马就说,对!
- [ ] 讲解辛普森悖论

- [ ] 分析快手的用户来源渠道
- [x] DAU下降5%怎么分析

> **参考**，分内因和外因

> 首先分明场景，看是否是活动日后，还是平常阶段的下降。
>
> 环比和同比看下降的速度，是突发的还是持续的
>
> 因为DAU是平常日更关心的指标，而转化率留存率是活动后才更关心的指标。

- [x] ！ MAU下降了，怎么分析
- [ ] ！ 现在一部分用户共有100w的总关注数，如何预测一年后的总关注数
- [x] ！ 搞促销活动，目的是提升销售额，怎么选有潜力的卖家进行合作
- [x] 选一个熟悉的APP

> 微信读书：

1. 阅读软件来说，阅读体验很好，书籍非常多，拥有很多经典的，口碑很好的专业书籍。而且用户可以通过登录和阅读来获取免费的阅读时长，可以满足日常阅读的需要。
2. 其次我非常喜欢它的批注功能，长按之后直接滑动就可以选择，而不像某些软件那样需要松开后才能选择，而且它的高亮笔刷做的很漂亮，不是简单的矩形，而是有起笔和停顿的。
3. 另外还有它的笔记分享功能，在设置里可以打开查看他人笔记的开关，这样在阅读时就可以看到其他用户批注次数较多的段落，尤其是在阅读一些技术类书籍时很有帮助，因为可以直接看到别人的疑问和讨论。
4. 除此之外，作为一个关注了许多技术类公众号而且经常写笔记的人来说，微信读书有一个公众号文章的剪藏功能做的很棒。只要是微信公众号里发布的文章都可以一键保存到微信读书里，然后做一些批注。

当然也是有一些缺点的，

1. 新出版的书籍比较少，这时就只能买纸质书了。
2. 另外，其实我更期望能够把微信读书作为我的笔记汇总软件，但目前只能剪藏公众号里的文章，不能处理其他的网页文章。

- [x] 抖音快手对比

功能设计：

- 首页设计

  抖音：大屏瀑布流式设计，上下滑动就可以直接播放。右下角有原声。朋友。切换大屏好像没有效果，或者说效果不明显，没有gif示例。

  快手：多屏+点击播放或大屏自动播放。同城。

- 搜索功能【对于我个人的使用来说，我其实不太喜欢刷视频的，而是更倾向于把这些APP当做工具，如搜索工具，所以我很关注搜索功能】
  - 搜索前
    - 快手：历史搜索为按钮，快手热榜，有一个新的“图片”
    - 抖音：历史搜索为长条，抖音热榜
  - 搜索结果【从上往下说】
    - 快手：
      - 关键词和常用筛选条件放一行了，
      - 如果可以，会显示“用户”，“直播”，“作品”等不同模块，还是挺好用
    - 抖音：
      - 关键词和筛选是分两行的
      - 搜索结果只直接显示作品，用户之类的还要左右滑动
- 评论功能

  快手：点击后视频仍在播放

  抖音：

- 用户主页
  - 快手把用户作品和APP通用设置分隔开了，抖音是传统的合并在一起的。
- 分享
  - 快手可以直接分享到微信，抖音需要下载而且只能复制链接。这也能解释为什么导航栏有区别了。

业务区别：

- 记录美好生活，拥抱每一种生活。快手北方人，抖音南方人？
- 抖音音乐作品多一点，快手？？？

- [ ] 应用设计：单列/双列设计的差异

> http://www.woshipm.com/pd/3304876.html

- [x] ！如何做一个能出圈的业务
  - [ ] 如果做一个业务，怎么验证**出圈**与否

    用abtest检测两种业务用户群体，看有无stat significant diff？？？
- [x] 漏斗分析怎么用

> 用于展示整个流程过程中转化率的变化，

- [x] 新生美妆up主投稿，怎么选择有潜力的进行培养？

> 答：
>
> ① 根据业务背景找到该情景下衡量新生美妆up主的北极星指标（e.g. 每周粉丝增长数，播放量等），作为因变量
> ② 结合视频类app的指标框架，可从内容观看、内容互动、粉丝转化、粉丝粘性等角度找到相应的指标，作为自变量
> ③ 进行预测，可以使用简单的线性回归，并结合常用的特征筛选方法进行特征筛选，优化模型
> ④ 最后剩下的指标构成的模型就是对潜力up主预测的合理模型

- [x] 数据分析最基础的思路是什么？

> 细分分析：互斥拆分，正交拆分

- [ ] 快手识别涉及赌博的风险用户，怎么构建特征？
- [x] 平常使用快手的路径
- [ ] 对快手有什么改进建议吗？
- [ ] 搜索功能在快手的定位

>

- [x] 如何判断用户/创作者是否有价值，要考虑哪些因素？

  【打算做活动，怎么根据数据筛选出有潜力的活动项目，怎么吸引用户？】

> 创作者分三种类型：
>
> UGC，用户生产内容，指一般用户

PGC，专业用户，拥有专业知识的，拥有一定权威的舆论领袖

OGC，两个主体：新媒体从业者，传媒行业人员创作；另一类是行业的精英，专业人士，与PGC一样

衡量创作者的价值，主要从内容质量和用户变现能力来衡量。

内容质量：

- 观看者数量，观看者页面停留时长，点赞率，视频完播率
- 评论数，转发数，代表了用户作品的传播能力，具有潜在的商业变现价值
- 转化率，如观看视频后由游客变为粉丝。用户粘性数据（重复活跃观看的用户数据），【如果创作者的作品缺少多样性，表现形式如不点击动态，只是点赞】
- 对创作者的粉丝进行分析，对用户画像进行描述，如用户质量高，付费行为多，也能说明创作者的价值较高

变现能力：

- 活动参与人数（点击率）
- 点击转化率（进行实际购买行为的）

- [ ] 衡量短视频的好坏

> 指标怎么加权

- [ ] 一般从什么角度进行数据分析？怎样的数据挖掘能真正对业务起到作用？
- [x] 瀑布流和双列点选的区别，从哪些角度进行分析？

> 基本体验
>
> 用户主动使用和被动使用。双列易于查找，单列易于提高使用时间

- [x] 广告投放的逻辑是什么？

> 比如微信朋友圈广告：
>
> - 微信朋友圈需要先发出需要广告的请求【投放广告的低价，如一条100元】
> - 第三方广告平台接收到需求后，在自己的广告库存中寻找满足要求的广告，从而填充这个请求【返回≥100元的广告】
> - 向用户展示广告

- [x] 广告收入的拆解方式

> 方法一：广告收入=（DAU*人均vv*ad load）/1000*CPM
>
> - DAU对广告收入的影响：
>   - 用户质量——广告变现收入的天花板
>   - 可展示广告DAU
> - 人均vv，人均Video View【核心指标】
> - DAU*人均vv反映的是一个用户侧的数据，决定了app商业化变现的量级
> - ad load，广告在信息流中的密度。比如刷多少条视频会遇到一个广告。是一个平衡用户侧和商业化侧的指标，如果用户对广告敏感的话，那么就需要降低ad load，从而提高用户体验。

- cpm，cost per mille，千人展示成本。不同广告主能够接受的cpm出价是不同的 。广告平台希望cpm报价越高越好，但不是所有的广告主都是一级广告主。因此为了满足用户多样化的需求，广告平台需要丰富自己的广告主结构。

方法二：广告收入=（请求总量×填充率×展现率）/1000×CPM

- 请求数量=可展示广告DAU×人均请求数量

  广告请求有两个规则：数量间隔，时间间隔。所以DAU越高，使用时长越高，则请求数量越高
- 填充率：是否有足够多的用户能够接收广告平台的库存。比如朋友圈为了提高填充率，**会进行差异化的低价策略。**
- 展现率：

方法三：广告收入=请求总量×填充率×展现率×点击报价×点击率【付费方式拆解】

- 点击率：广告主素材的内容质量、平台推荐系统能力、CTR预估的准确性

- [x] 给广告设置埋点

如果点击，点击广告时，记录bannerid和userid，以及点击广告前最后一个操作到点击广告的时间间隔。

如果关闭，看加载完成到关闭广告之间的时间，关闭广告之前的最后一个行为

- [ ] 数据埋点的流程：

数据埋点一般时伴随新的业务/功能产生的，因为这是一个多部门复合型的工作。

埋点设计：

- 产品经理明确新业务的需求、UI设计图，后续关注的一些业务点，整理成文档，数据分析师完成**埋点设计**，埋点文档反馈给产品经理和开发，最后结果会返回给数据分析师，用于**埋点验证**。

埋点验证：

- 明确业务背景【比如产品经理要策划训练营活动，搞清楚训练营有哪些活动入口，用户进入后会经历哪些环节，是不是存在漏斗转化，活动本身有哪些需要关注的指标，活动最终会作用于哪些北极星指标】就决定了设计埋点时应该往哪些方向考虑。埋点不能太多，不能太少。
- 规范埋点的命名，如图所示

事件	事件名	埋点性质	参数值	备注
用户登录	user_log_in	新增埋点	-user_id -enter_from (appA_loggin_page, A产品登录界面 appB_loggin_page, B产品合作登录界面) -timestamp	用户在登录时触发

不同系统的埋点开发不同

埋点验收：

- 需要触发埋点的环节，所有的埋点是否都能正常上报
- 上报的埋点中，是否包括需要的参数值，拼写是否正确

- [ ] 预测广告费用：时间序列模型

- [ ] 快手电商应该关注哪些指标?

> 需求侧：GMV，转化率，退单率
>
> 供给侧：？？？

- [ ] 如果做一个看板，会放什么指标上去?

项目

- [ ] 知道什么分类算法?

> 逻辑回归，决策树，朴素贝叶斯，SVM，一些神经网络模型如简单的CNN。比赛用的lgbm，

  - [ ] lgbm的优缺点?
  - [ ] 简单介绍lightgbm
  - [ ] 口述SVM

- [x] 决策树详细
  - [x] 特征分割用什么
  - [x] 决策树的优点，缺点，如何规避
  - [x] 样本不平衡怎么解决
- [x] 逻辑回归和线性回归的区别

> 回归和分类，损失函数，评价指标

- [x] 损失函数有了解吗?

> 损失函数的对比：
>
> - 对数损失函数和均方差损失的区别：求导后发现，梯度下降速度不同
>
> 不同模型的：
>
> - 线性回归，均方误差
> - 逻辑回归：对数损失
> - SVM：合页损失
> -

- [x] 对特征工程的理解
- [x] 类别特征，数值特征的处理
- [x] SVM原理
- [ ] 模型用来做什么？遇到的问题？**怎么改进**？
- [x] 逻辑回归的损失函数，kmeans
- [ ] 项目结果怎么落地的?
- [x] 模型评价指标

> 分类：ACC，P，R，AUC，F1 score，
>
> 回归：MSE，

- [ ] ROC和PR曲线的区别，形状区别？什么时候选PR，什么时候选ROC？

- [x] 朴素贝叶斯的理解

> A：基于朴素贝叶斯的分类模型，朴素是指“假设各个特征之间相互独立，不会互相影响，即条件独立性”。贝叶斯定理是基于假设的先验概率，给定假设下观察到不同数据的概率，以此计算后验概率

- [ ] ARMA模型
- [x] 对text minning 的理解？

> 从大量的文本中筛选出有价值的信息

- [ ] 怎么做文本向量化？

>

- [x] 样本不平衡怎么办？具体的采样方法还记得吗？

> 上采样，下采样，SMOTE，回译

- [ ] 快手某天违规率突然上升了10%，怎么分析
- [ ] 如何判断用户
- [x] lasso和岭回归

> lasso是绝对值正则项，
> 岭回归是平方正则项，

- [x] 随机森林

> 随机的意义：
> 随机抽一部分样本用于分支；随机选一部分特征作为特征子集用于分支。
> 不易过拟合

- [x] 决策树
  - [ ] 怎么分支

    > 信息增益，选取信息增益最大的特征

  尽管剪枝了，仍易过拟合
- [ ] 针对某个产品/业务，搭建相应的指标体系
- [ ] 解释什么是神经网络

> wx+b，更新权重，反向传播？？？

- [ ] 数据处理、one-hot、标准化等等操作的函数写下来,然后边写边讲解

统计

- [x] 假设检验
- [x] 极大似然估计
- [ ] 第一类和第二类哪个重要
- [ ] 标准差和标准误

> https://www.zhihu.com/question/22864111

SQL

- [x] 每个用户得分最高的视频，
- [ ] 开播三分钟内无人进入的房间号
- [x] 主播id，主播类型，主播粉丝数，求各个类型主播粉丝数top100的主播
- [x] SQL语句执行顺序

> FROM，JOIN ON，WHERE，GROUP BY，HAVING，SELECT，ORDER BY

- [x] video_table：video_id,user_id。user_table：user_id,age,city。年龄20以下的用户，每个city随机抽样100数据
- [x] user_id,time,action，用户在点击行为之前进行的最多的行为（曝光除外）

```
select count(distinct action) as num
from
left join tb as b
on a.user_id=b.user_id
and time<(select min(date) from tb
          where action='click'
          group by user_id)
```

- [ ] 用户点击行为表：is_click=1：user_id,date,is_click。用户日活：date,user_id。
    - 某天有点击和没点击的用户总数?

```
  select is_click,count(distinct user_id)
  from
```

    - 对某天有点击和没点击的用户，分别求第二天的留存率

```
select first_day,count(distinct b.user_id) as sec_num
from(select user_id,min(date) as first_day
    from B
    group by user_id) a
left join B b on a.user_id=b.user_id and DATEDIFF(a.first_day,b.date)=1
group by a.first_day
```

- [ ] lag() over() 与 lead() over()

> https://xiaoshuwen.blog.csdn.net/article/details/107188400

- [ ] SQL调优的经验，大数据上提高效率。

> 使用join的时候，先把row drop掉，然后再合并。
>
> - 把用不到的行和列先排除
> - 左表里的每一行都在右表里找，因此可以把小表放在左边

- [x] 每个班级排名前十的同学，上学期到本学期进步最大的同学和分数
- [ ] group by的key中有null,会怎么样
- [x] 日期，关注着ID，被关注着ID，求每一天的双关数

> 自联结，表1的关注者=表2的被关注者，count，group by 日期

  - [ ] 主播id，观看者id，互相看过对方视频的用户

- [ ] 留存率计算：user_id,p_date

> https://blog.csdn.net/tsyh8797/article/details/103597215?utm_medium=distribute.pc_relevant.none-task-blog-BlogCommendFromMachineLearnPai2-1.channel_param&depth_1-utm_source=distribute.pc_relevant.none-task-blog-BlogCommendFromMachineLearnPai2-1.channel_param

- [ ] 每个用户最长连续登录天数

> 自连接，用最大的减去最小的？？？

- [ ] 8月5日之后没有登录过的用户
- [x] 字符串分割

substring方法：

```
select substring(name,3,4) --从3开始，长度为4【下标为1】
from student;
```

substring index方法：

```
-- https://blog.csdn.net/weixin_38929027/article/details/106688308
```

- [x] 求互关用户之间的关系链长度【最短路径，有环】

```
-- 创建数据
DROP TABLE IF EXISTS `abr`;
CREATE TABLE `abr`  (
  `aid` int(0) NULL DEFAULT NULL,
  `bid` int(0) NULL DEFAULT NULL,
  `distance` double NULL DEFAULT NULL
) ENGINE = InnoDB CHARACTER SET = utf8mb4 COLLATE = utf8mb4_0900_ai_ci ROW_FORMAT
= Dynamic;

-- ----------------------------
-- Records of abr
-- ----------------------------
INSERT INTO `abr` VALUES (1, 2, 1);
INSERT INTO `abr` VALUES (1, 6, 1);
INSERT INTO `abr` VALUES (1, 3, 1);
INSERT INTO `abr` VALUES (1, 4, 1);
```

```
INSERT INTO `abr` VALUES (1, 5, 1);
INSERT INTO `abr` VALUES (2, 1, 1);
INSERT INTO `abr` VALUES (2, 3, 1);
INSERT INTO `abr` VALUES (2, 4, 1);
INSERT INTO `abr` VALUES (3, 1, 1);
INSERT INTO `abr` VALUES (3, 2, 1);
INSERT INTO `abr` VALUES (3, 7, 1);
```

上述表的关系链为：【补一张图】

用递归方法求最短路径，参考大佬文章**SQL求最短路径**，【面试中题干和表结构没有表示清楚，所以这里加了一个distance为1，替代ABBA表示长度，便于求最短】【补一些中间结果的图，如**这个**】

```
with recursive t as
(
    select *,cast(concat(a.aid,'>',a.bid) as char(100)) as path
        from abr a
        where aid = 1 --此时这条select语句的结果已经存到临时表t里面了
    union all
    select
        t.aid,b.bid, -- 用于拼接两段路径
        t.distance+b.distance,
        cast(concat(t.path,'>',b.bid) as char(100)) as path
        from t
    inner join abr b
        on t.bid=b.aid -- t的结束是b的开始，如t.bid=2，b.aid=2，那就说明2是两个路径的交点
        and instr(t.path,b.bid)<=0 -- 避免因为环导致的死循环，即已在路径里不能再用于计算
),
t1 as -- 表t1用于从相同start和end的不同路径中选取最短的
(
    select *,row_number () over (partition by aid,bid order by distance) as rn
    from t
)

select aid,bid,distance as min_distance
from t1
where bid=7 and rn=1; --添加关系链的终点，即结点7
```

规划

- [x] 为什么投数据分析岗?
- [ ] 对自己的分析技能怎么评价？有什么强的、弱的?

> 答：SQL/机器学习有经验，互联网了解多吗?

- [ ] 偏ds还是da

> 答：偏da，因为自己缺少这方面的实践。如果部门结构安排，做ds也可以接受

- [ ] 假设入职，首先希望学到什么技能?
- [ ] 为什么选择你？而不选择其他人?

- [ ] 为什么不选你?

- [ ] 优点和缺点?

> 缺点：~~缺少实习，业务实践?~~，面试官反馈这应该不是一个缺点？很滑头的一个说法？
>
> 数字敏感性不够，但之后可以靠经验弥补
>
> 优点：学技术的，技术支持没问题

- [ ] 数据分析不同岗位的理解?
- [x] 被夸还是被骂进步的快?

> 被夸会进步，但是可能没有被骂进步的那么快。因为被骂是说明我有了明显的问题，但是被夸的时候，虽然我会有满足感，也会做一些努力继续维持这种会被夸的状态，但这样我是不知道自己有什么问题的，也缺少更深入研究的动力。小结一下就是都会进步，但被骂进步更快。

- [x] 身边不可超越的人

> 室友唐百川，室友孙逸文。
>
> 怎么克服的？唐百川技术层面：前期是他带着我学的，后来入门之后我就会自己找一些学习资料，当然肯定不会私藏啊，我们的微信聊天里面都是发的技术文章。超越算不上，但如果在具体某个事件里超越，那是有可能的，比如一些课的成绩，但总之还是不容易赶上的。

- [ ] 三个词形容自己，举例子
- [ ] 自己做数据分析的优势

反问

- [x] 团队构成
- [ ]

基本概念

- [ ] **指标大全-产品设计师要了解的数据指标**，
- [ ] 用户生命周期价值LTV，lifetime value：第一次接触产品到最终离开成为流失用户之内所创造的价值（净利润的估计值，不是GMV，需要扣除广告投入等）

  $LTV = LT * ARPU$，LTV类似于正态分布的曲线

  - [ ] 延伸LT
  - [ ] 预测LTV（预测结果会指导是否要获取新用户）

> - 方法一：LTV公式计算。【简单常用，缺点：留存率预测误差，ARPU动态变化】
>
>   LT=1+次日留存率+2日留存率+...+n日留存率，对LT的预测，其实就是对留存率衰减的预测。由于不同周期内的ARPU不稳定，选取一定时期内ARPU的均值。
> - 方法二，交易成交角度预测：LTV=付费用户LT*ARPU*付费转化率，同样要预测留存率和ARPU，更适合游戏、电商方向的LTV计算。

> - 方法三，LTV时间序列，预测同样人群的LTV，天数越多，则精度越高。

- [ ] ARPU，average revenue per user
- [ ] CAC，customer acquisition cost，用户增长时的一个原则：CAC要小于LTV。