

1. 如何利用电商漏斗减少流失率？[业务]

(1) 分解漏水过程：

弄清楚用户从进入网站到购买经历了几个步骤。一般来说用户会经历浏览首页-中间页（分类页、搜索页）-产品页-购物车-结算等几个步骤。接着弄明白每个环节流失了多少用户。



(2) 排查每个环节的漏洞

按照漏水的顺序，一个环节一个环节分析下去

a. 首页弹出率分析：

每天来的新客户占多少？老客户占多少？新老客户的弹出率分别是多少？

新老用户的弹出率容易考验网站的基础能力，新客户的弹出率可以检验一个网站抢客户的能力。对于老客户来说，流程上的用户体验相对不是最讲究，产品的质量和价格是吸引老客户的关键。

一般的来说，如果是一个新网站（店铺），拓展新用户比经营老客户更为重要的话，新老客户的比例最好是在 6:4 (甚至 7:3)，那么首页就要有一些手段偏向抓住新客户。如果新用户的弹出率非常高，或者是老用户的弹出率非常高，那么运营者就该反思，是不是网站（店铺）首页的设计没有照顾到新客户或者老客户。

b. 渠道流量弹出率分析

流量分几个大渠道进来，每个渠道的弹出率情况如何？

流量渠道有多种，主要有付费渠道和免费渠道，每类渠道又可细分为多条路径，每条路径进来的流量弹出率可能差异非常大。同时，渠道流量的着陆页也会不同，着陆页的弹出率也会不尽相同。针对自己的主要流量渠道排查下去，很容易发现，哪条渠道在漏水。找到了痛处之后，再找到相应的解决方法就不难了。

c. 首页被点击最多、最少的地方是否有异常情况？

在首页，点击次数异常高或者异常低的地方，应该引起注意。

一般来说，首页的“E”（以E字中间的“一”为界，上部是首页第一屏）部份是最抓用户眼球的地方，在这个“E”上如果出现点击次数较低的情况，就属于异常情况，应当注意，或者干脆移到“E”外面去；同理，如果在“E”的空白处出现了点击次数较高的情况，也可分析原因，可考虑要不要移到“E”上面来。一般商城

(店铺) 首页非常长, 许多用户不会浏览到首页底部, 所以“E”最下面的“一”就往往可去掉, 变成了“F”规律。

d. 中间页分析

中间页一般包括产品目录页、促销页、搜索页。

怎么判断促销、目录和搜索是否成功, 就看一下走到产品页的用户百分比是多少, 哪一个渠道走得不好, 就要改善。

e. 产品页要特别留意用户停留时间

到了产品页, 用户留不留, 与产品描述、质量有非常大的关系。所以, 要特别留心客户停留在产品页的时间, 如果许多用户打开产品页不到 3 秒钟就走了, 就要留意分析原因了。是不是这个产品没有吸引力? 是不是产品描述不准确?

f. 购物车里多少产品没有付款?

许多用户把产品放进购物车, 但是并不付款。这时候, 要多思考为什么这么多用户放在购物车里却不付款? 如果找不到用户不付款的原因, 可以直接向用户进行电话访问。也可分析同时被放在购物车的产品之间关联性。

2. 估算北京有多少加油站?

解析:

角度一: 北京加油站数量=北京一天需要加油的车的数量÷每天每个加油站的容量
先来算北京市一天需要加油的车的数量。北京约有 2000 万人口, 按 3 人为一户来说有 667 万户人, 假设一户人有一辆车, 则共有车 667 万辆, 若一辆车 10 天加一次油, 则一天需要加油的车辆为 66.7 万辆。

接着算每天每个加油站的容量。一个典型的加油站有 3 杆加油枪, 加一辆车需要 5 分钟。在高峰期(按 6 小时算), 一小时一杆加油枪可以加 $60 \div 5 = 12$ 辆车, 一个加油站可以加 36 辆车。在非高峰期, 按 5 分钟来一辆车, 则一小时加 12 辆车。那么, 每天每个加油站的容量=高峰期加的车数量+非高峰期加的车数量
 $= 36 \times 6 + 12 \times 18 = 432$ (辆)

因此, 北京加油站数量为: $667000 \div 432 = 1543$ (座)

角度二: 北京加油站数量=北京市面积*加油站分布密度

北京市面积 16410 平方千米。按照《城市道路交通规划设计规范》的要求, 城市加油站的服务半径 $R = 0.9 \sim 1.2$ km, 服务面积 $A = \pi R^2 = 2.54 \sim 4.52$ 平方公里, 折算分布密度 $\rho = 0.22 \sim 0.39$ 座/平方公里。已知北京五环内城市发展更快, 人口车辆更为集中, 加油站密度更高。北京市五环内面积为 735 平方千米, 加油站数量约为 735×0.39 约等于 287 座; 五环外面积为 15700 平方千米, 加油站数量为 $15700 \times 0.22 = 3454$ 座

因此, 北京加油站数量为: $287 + 3454 = 3741$ (座)

结果分析: 根据两个角度的估算, 可知北京加油站数量约在 1000~4000 区间。仍有一些因素可能导致误差, 如北京人口中外来务工人口较多, 许多人未购车, 会影响北京市车辆总数这一因素的准确性。

3. x^x 如何求导 [数学基本功]

1) 对数求导法

$$y = x^x$$

$$\ln y = x \ln x$$

$$\frac{y'}{y} = \ln x + 1$$

$$y' = x^x (\ln x + 1)$$

($x^x > 0$)

2) 指数复合求导

$$(x^x)' = (e^{x \ln x})' = x^x (\ln x + 1)$$

3) 复合求导

$$y = u(s, t) = s^t$$

$$s = f(x); t = g(x)$$

$$f(x) = g(x) = x$$

按链式法则展开

$$\frac{dz}{dt} = \frac{\partial z}{\partial x} \frac{dx}{dt} + \frac{\partial z}{\partial y} \frac{dy}{dt}$$

$$\begin{aligned}y' &= ts^{t-1}f'(x) + s^t \ln sg'(x) \\&= g(x)[f(x)]^{g(x)-1}f'(x) + [f(x)]^{g(x)} \ln f(x) \cdot g'(x) \\&= f(x)^{g(x)} \left[\frac{g(x)}{f(x)} f'(x) + g'(x) \ln f(x) \right] \\&= x^x (1 + \ln(x))\end{aligned}$$

4) 按定义展开

$$\begin{aligned}(x^x)' &= \lim_{h \rightarrow 0} \frac{(x+h)^{x+h} - x^x}{h} \\&= \lim_{h \rightarrow 0} \frac{(x+h)^{x+h} - (x+h)^x + (x+h)^x - x^x}{h} \\&= \lim_{h \rightarrow 0} (x+h)^x \lim_{h \rightarrow 0} \frac{(x+h)^h - 1}{h} + x^x \lim_{h \rightarrow 0} \frac{\left(1 + \frac{h}{x}\right)^x - 1}{h} \\&= \lim_{h \rightarrow 0} (x+h)^x \lim_{h \rightarrow 0} \frac{(x+h)^h - 1}{h} + x^x \lim_{h \rightarrow 0} \frac{\left(1 + \frac{h}{x}\right)^{\frac{x}{h}h} - 1}{h} \\&= x^x \ln(x) + x^x \cdot \ln(e) \\&= x^x (\ln x + 1)\end{aligned}$$

4. S 市 A, B 共有两个区, 人口比例为 3: 5, 据历史统计 A 的犯罪率为 0.01%, B 区为 0.015%, 现有一起新案件发生在 S 市, 那么案件发生在 A 区的可能性有多大? [概统]

- A. 37.5%
B. 32.5%
C. 28.6%
D. 26.1%

答案: C

解析:

在 A 区犯案概率: $P(C|A)=0.01\%$

在 B 区犯案概率: $P(C|B)=0.015\%$

在 A 区概率: $P(A)=3/8$

在 B 区概率: $P(B)=5/8$

犯案概率: $P(C)=(3/8*0.01\% + 5/8*0.015\%)$

则犯案且在 A 区的概率: $P(A|C)=P(C|A)*P(A)/P(C)=0.01\%*(3/8)/(3/8*0.01\% + 5/8*0.015\%) \approx 28.6\%$

5. 一个包里有 5 个黑球, 10 个红球和 17 个白球。每次可以从中取两个球出来, 放置在外面。那么至少取_____次以后, 一定出现过取出一对颜色一样的球。 [概统]

- A. 16
- B. 9
- C. 4
- D. 1

答案: A

解析: 考虑最坏的情况, 前 10 次取出的都是红球+白球的组合, 后 5 次取出的都是黑球+白球的组合, 最后只剩下两个白球, 则再取 1 次必取出相同颜色的球, 因此总计 16 次。

6~7 题 [SQL]

表 user_id, visit_date, page_name, plat

- 6. 统计近 7 天每天到访的新用户数
- 7. 统计每个访问渠道 7 天前的新用户的 3 日留存率和 7 日留存率

解析:

1) 近 7 天每天到访的新用户数

```
select day(visit_date), count(distinct user_id)
from table
where user_id not in
(select user_id from table
where day(visit_date) < date_sub(visit_date, interval 7day))
```

2) 每个渠道 7 天前用户的 3 日留存和 7 日留存

三日留存

先计算每个平台 7 日前的新用户数量

```
select t1.plat, t1.c/t2.c as retention_3
(select plat, count(distinct user_id)
from table
group by plat, user_id
having day(min(visit_date)) = date_sub(now(), interval 7 day)) as t1
left join
(select plat, count(distinct user_id) as c
from table
group by user_id having count(user_id) > 0
having day(min(visit_date)) = date_sub(now(), interval 7 day)
and day(max(visit_date)) > date_sub(now(), interval 7 day)
and day(max(visit_date)) <= date_sub(now(), interval 4day)) as t2
on t1.plat = t2.plat
```

8~10 题[SQL]

有 3 个表 S, C, SC:

S (SNO, SNAME) 代表 (学号, 姓名)

C (CNO, CNAME, CTEACHER) 代表 (课号, 课名, 教师)

SC (SNO, CNO, SCGRADE) 代表 (学号, 课号, 成绩)

问题:

8. 找出没选过“黎明”老师的所有学生姓名。
9. 列出 2 门以上 (含 2 门) 不及格学生姓名及平均成绩。
10. 既学过 1 号课程又学过 2 号课所有学生的姓名。

解析:

8. — 考察条件筛选

```
select sname from s where sno not in
( select sno from sc where cno in
  (
    select distinct cno from c where cteacher='黎明'
  )
);
```

9. — 考察聚合函数, 条件筛选

```
select s.sname, avg_grade from s
join
(select sno from sc where scgrade < 60 group by sno having count(*) >=
2) t1
on s.sno = t1.sno
join
(select sno, avg(scgrade) as avg_grade from sc group by sno ) t2
on s.sno = t2.sno;
```

10. — 考察筛选、连接

```
select sname from
( select sno from sc where cno = 1) a
join
(select sno from sc where cno = 2) b
on a.sno = b.sno
```

11. 一般, K-NN 最近邻方法在 () 的情况下效果较好 [算法]

- A. 样本较多但典型性不好
- B. 样本较少但典型性好
- C. 样本呈团状分布
- D. 样本呈链状分布

答案: B

解析：样本呈团状颇有迷惑性，这里应该指的是整个样本都是呈团状分布，这样 kNN 就发挥不出其求近邻的优势了，整体样本应该具有典型性好，样本较少，比较适宜。

12. 下列不是 SVM 核函数的是：

- A. 多项式核函数
- B. logistic 核函数
- C. 径向基核函数
- D. Sigmoid 核函数

答案：B

解析：

SVM 核函数包括线性核函数、多项式核函数、径向基核函数、高斯核函数、幂指数核函数、拉普拉斯核函数、ANOVA 核函数、二次有理核函数、多元二次核函数、逆多元二次核函数以及 Sigmoid 核函数

13. (多选)数据清理中，处理缺失值的方法是？[算法]

- A. 估算
- B. 整例删除
- C. 变量删除
- D. 成对删除

答案：A, B, C, D

解析：

数据清理中，处理缺失值的方法有两种：

删除法：

- 1) 删除观察样本
 - 2) 删除变量：当某个变量缺失值较多且对研究目标影响不大时，可以将整个变量整体删除
 - 3) 使用完整原始数据分析：当数据存在较多缺失而其原始数据完整时，可以使用原始数据替代现有数据进行分析
 - 4) 改变权重：当删除缺失数据会改变数据结构时，通过对完整数据按照不同的权重进行加权，可以降低删除缺失数据带来的偏差
- 查补法：均值插补、回归插补、抽样填补等成对删除与改变权重为一类估算与查补法为一类

14. 小易有一个长度为 n 的整数序列 a_1, \dots, a_n 。然后考虑在一个空序列 b 上进行 n 次以下操作：[Python]

- 1、将 a_i 放入 b 序列的末尾
- 2、逆置 b 序列

小易需要你计算输出操作 n 次之后的 b 序列。

解析：

在这里一定不要被题迷惑，其实不需要逆序，寻找规律

规律题

$n = 1, b = 1$ $n = 1$ 直接输出

$n = 2, b = 2, 1$

$n = 3, b = 3, 1, 2$

$n = 4, b = 4, 2, 1, 3$

$n = 5, b = 5, 3, 1, 2, 4$

$n = 6, b = 6, 4, 2, 1, 3, 5$

由上述可推，当 n 为奇数时，

先从后向前输出奇数位置的数字，再从前向后输出偶数位置的数字

当 n 为偶数时

先从后向前输出偶数位置的数字，再从前向后输出奇数位置的数字

```
17 n = int(input())
18 num = [int(x) for x in input().split()]
19 def findNum(num, n):
20     if n == 1:
21         print(num[0])
22     if n % 2 == 0:
23         for i in range(n-1, 0, -2):
24             print(num[i], end=" ")
25         for i in range(0, n-2, 2):
26             print(num[i], end=" ")
27
28         print(num[n-2], end=" ")
29     else: # n%2 == 1
30
31         for i in range(n-1, 0, -2):
32             print(num[i], end=" ")
33         for i in range(0, n-2, 2):
34             print(num[i], end=" ")
35 findNum(num, n)
36
```

4 输入
1 2 3 4
4 2 1 3 输出

15. 运行以下 Python 表达式后，X 的值为？ [Python]

$X = 3 == 3, 5$

A, 3

B, 5

C, (True, 5)

D, (False, 5)

答案：C

解析：

该题考察了对 Python 中赋值及表达式的运用。 $x=3$ 先进行赋值，再进行比较后得到 True，后边的逗号使用该句的返回值变成一个 tuple，5 为普通数字。