1. 如果次日用户留存率下降了 5%该怎么分析［业务］

这道业务题考的是大家的业务嗅觉和相关模型的运营~答案属于开放式的，大家言之有理即可，下面小学生提供几个思路供大家参考~

答案：先判断这个下降是否合理，然后从各个方向头脑风暴

（1）首先采用“两层模型”分析：对用户进行细分，包括新老、渠道、活动、画像等多个维度，然后分别计算每个维度下不同用户的次日留存率。通过这种方法定位到导致留存率下降的用户群体是谁。

a. 按照获取客户渠道进行分析

b. 按照获取客户时间进行分析

c. 按照用户行为进行分析

d. 不同群组对产品不同模块使用状况的分析

（2）分析留下来用户的核心需求和流失用户的流失原因

内部因素：

分为获客（渠道质量低、活动获取非目标用户）、满足需求（新功能改动引发某类用户不满）、提活手段（签到等提活手段没打成目标、产品自然使用周期低导致上次获得的大量用户短期内不需要再使用等）；

外部因素：

采用 PEST 分析，政治（政策影响）、经济（短期内主要是竞争环境，如对竞争对手的活动）、社会（舆论压力、用户生活方式变化、消费心理变化、价值观变化等偏好变化）、技术（创新解决方案的出现、分销渠道变化等）

同类题：

如果某天的某电商 app 的 DAU 下降了很多，你怎么分析？

某电商什么时候 DAU 最高？除了活动日，平常日呢？

这两道题与上题答题思路相似。

2. 如何估算中国 K12 课外英语辅导的市场［估算］

解析：

估算题是经常出现在咨询等岗位面试中的题目，体现的是对面试者思维、逻辑、拆解问题能力的考察，具体的背景介绍在 Day1 已经说过，小学生就不啰嗦啦！

此题可从需求角度与供给角度来解答

角度一（需求层面）：

中国 K12 课外英语学员=中国处于 K12 阶段人群数量*参与英语培训的人数*英语培训年花费

中国总人口数约 14 亿，K12 年龄段人口比例约 17%，则中国处于 K12 阶段的人群数量=14 亿*17%=2.38 亿。

K12 教育在今年不断升温，参与课外培训的学生约有 50%，其中参与了英语课外培训的学生约有 20%，则参与英语培训的人数=2.38 亿*50%*20%=0.238 亿。若英语培训年花费约为 1 万，则

中国 K12 课外英语辅导市场=0.2856 亿*1 万=2380 亿

角度二（供给层面）：
中国 K12 课外英语学员=中国 K12 阶段所有教师数*参与课外英语辅导教师占比*课外英语辅导老师的平均工资
根据人口统计信息，可以得知小学阶段公立教师约有 780 万，初中阶段公立教师约有 495 万，高中阶段公立教师约有 330 万，则中国 K12 阶段的公立教师人数=780+495+330=1605 万。
若其中英语老师占比 20%，参与课外英语辅导的教师占比 20%，则参与课外英语辅导的公立教师人数=1605 万*20%*20%=64.2 万。根据常识，不仅仅有在校的公立教师，还有一部分私立学校、课外培训机构的教师，非公立教师人数占比约为公立教师的 80%，则参与课外英语辅导的私立教师人数=64.2 万*80%=51.36 万。则中国 K12 阶段所有参与课外英语辅导的教师数量=64.2 万+51.36 万=115.56 万。
根据市场信息，课外英语辅导平均月收入约 8000 元，则课外英语辅导年平均收入=8000 元*12=9.6 万。
由此，中国 K12 课外英语辅导市场= 115.56 万*9.6 万=1109.376 亿

结果分析：根据两个角度的估算，中国 K12 课外英语辅导的市场约为 1000~3000 亿元。仍有一些因素可能导致误差，如，有意愿参加课外培训的学生比例尚待考证，可通过市场调查使其更加精确。

3. 以下属于避免决策树过拟合的方法是（可多选）[算法]

A. 限制树深
B. 剪枝
C. 限制叶节点数量
D. 数据增强（加入有杂质的数据）

答案：ABCD
解析：
在这里小学生给大家补充一下产生过度拟合数据问题的常见原因有哪些
原因 1：样本问题
（1）样本里的噪音数据干扰过大，大到模型过分记住了噪音特征，反而忽略了真实的输入输出间的关系；（什么是噪音数据？）
（2）样本抽取错误，包括（但不限于）样本数量太少，抽样方法错误，抽样时没有足够正确考虑业务场景或业务特点，等等导致抽出的样本数据不能有效足够代表业务逻辑或业务场景；
（3）建模时使用了样本中太多无关的输入变量。
针对原因 1 的解决方法：
合理、有效地抽样，用相对能够反映业务逻辑的训练集去产生决策树；

原因 2：构建决策树的方法问题
在决策树模型搭建中，我们使用的算法对于决策树的生长没有合理的限制和修剪的话，决策树的自由生长有可能每片叶子里只包含单纯的事件数据或非事件数据，可以想象，这种决策树当然可以完美匹配（拟合）训练数据，但是一旦应用到新的业务真实数据时，效果是一塌糊涂。
针对原因 2 的解决方法（主要）：
剪枝：提前停止树的增长或者对已经生成的树按照一定的规则进行后剪枝。

上面的原因都是现象，但是其本质只有一个，那就是“业务逻辑理解错误造成的”，无论是抽样，还是噪音，还是决策树等等，如果我们对于业务背景和业务知识非常了解，非常透彻的话，一定是可以避免绝大多数过拟合现象产生的。因为在模型从确定需求，到思路讨论，到搭建，到业务应用验证，各个环节都是可以用业务敏感来防止过拟合于未然的

想要了解更多的同学可以参考这个链接：

https://blog.csdn.net/sinat_32043495/article/details/78729610

4. **请说出两种主要的降维方法［算法］**

解析：

方法一：

主成分分析（PCA）

将样本投影到某一维上，新的坐标的选择方式：找到第一个坐标，数据集在该坐标的方差最大（方差最大也就是我们在这个数据维度上能更好地区分不同类型的数据），然后找到第二个坐标，该坐标与原来的坐标正交。该过程会一直的重复，直到新坐标的数目和原来的特征个数相同，这时候我们会发现数据的大部分方差都在前面几个坐标上表示，这些新的维度就是我们所说的主成分。

（1）PCA 的基本思想：寻找数据的主轴方向，由主轴构成一个新的坐标系，这里的维数可以比原维数低，然后数据由原坐标系向新坐标系投影，这个投影的过程就是降维的过程。

（2）PCA 算法的过程

①将原始数据中的每一个样本都用向量表示，把所有样本组合起来构成样本矩阵，通常对样本矩阵进行中心化处理，得到中心化样本矩阵。

②求中心化后的样本矩阵的协方差；

③求协方差矩阵的特征值和特征向量；

④将求出的特征值按从大到小的顺序排列，并将其对应的特征向量按照此顺序组合成一个映射矩阵，根据指定的 PCA 保留的特征个数取出映射矩阵的前 n 行或者前 n 列作为最终的映射矩阵；

⑤用映射矩阵对数据进行映射，达到数据降维的目的。

方法二：

LDA（Linear discriminant analysis）

线性判别式分析，也叫 fisher 线性判别，是模式识别中的经典算法。

是一种监督学习的降维技术，它的数据集的每个样本是有类别输出的。

思想：投影后类内距离最小，类间距离最大。

1、线性判别：将高维的模式样本投影到最佳鉴别矢量空间，以达到抽取分类信息和压缩特征空间维数的效果，投影后保证模式样本在新的子空间有最大的类间距离和最小的类内距离，这是一种有效的特征提取方法，使用这个方法，能使得投影后模式样本的类间散布矩阵最大，且同时类内散布矩阵最小。

2、与 PCA 相比较：

（1）共同点：

①都属于线性方法；

②在降维时都采用矩阵分解的方法；

③都假设数据符合高斯分布；

（2）不同点：

①LDA 是有监督的；

②不能保证投影到的坐标系是正交的（根据类别的标注，关注分类能力）；

③降维直接与类别的个数有关，与数据本身的维度无关（原始数据是 n 维的，有 c 个类别，降维后一般是到 c-1 维）

④可以用于降维，还可用于分类；

⑤选择分类性能最好的投影方向。

5. **请说明随机森林较一般决策树稳定的几点原因［算法］**

- bagging 的方法，多个树投票提高泛化能力
- bagging 中引入随机（参数、样本、特征、空间映射），避免单棵树的过拟合，提高整体泛化能力

6. **假设检验的意义和应用［概统］**

答案：假设检验是推论统计中用于检验统计假设的一种方法。其基本原理是先对总体的特征作出某种假设，然后通过抽样研究的统计推理，对此假设应该被拒绝还是接受作出推断。

检验过程是比较样本观察结果与总体假设的差异。差异显著，超过了临界点，拒绝 H0；反之，差异不显著，接受 H0。

7. **方差如何计算？（可举例子说明）［概统］**

举个栗子，一组数据：48， 49，50，53，55，均值为：51

则方差为：

$$\frac{(48-51)^2+(49-51)^2+(50-51)^2+(53-51)^2+(55-51)^2}{5}$$

计算公式：

$$s_N^2=\frac{1}{N}\sum_{i=1}^{N}(x_i-\overline{x})^2$$

8. **设随机变量满足:E(X)=μ,D(X)=σ²,则由切比雪夫不等式，求{|X-μ|>=4σ}<=__.［概统］**

答案：1/16

解析：本题考查切比雪夫不等式（Chebyshev）的应用：

$$P(|X-E(X)|\geqslant b)\leq\frac{Var(X)}{b^2}$$

带入公式可得答案为：1/16

9. **如何利用 SciKit 包训练一个简单的线性回归模型 [Python]**
   解：
   利用 linear_model.LinearRegression()函数
   Lr_model = linear_model.LinearRegression()
   Lr_model.fit(X_train,y_train)

10. **如何在 python 中复制对象 [Python]**
    答案：使用 copy 包的 copy 和 deepcopy 函数。其中，copy 仅拷贝对象本身，而不拷贝对象中引用的其它对象；deepcopy 除拷贝对象本身，而且拷贝对象中引用的其它对象。

11. **例举几个常用的 python 分析数据包及其作用[Python]**
    数据处理和分析：NumPy, SciPy, Pandas
    机器学习：SciKit
    可视化： Matplotlib, Seaborn

12. **如何对 list 中的 item 进行随机重排 [Python]**
    答案：使用 shuffle()函数

13. **sql 中 null 与 ‘’ 的区别 [SQL]**

    答案：

- null 表示空，用 is null 判断
- ''表示空字符串，用=''判断

14. **SQL 中包含了哪些数据类型** [SQL]

    解析：不同的 SQL 里的名称除了个别可能会有差异但是大体上是一致的，总结来说主要有一下几种~

- 字符串：char、varchar、text
- 二进制串：binary、varbinary
- 布尔类型：boolean
- 数值类型：integer、smallint、bigint、decimal、numeric、float、real、double
- 时间类型：date、time、timestamp、interval

15. **现有一个数据库表 Tourists，记录了某个景点 7 月份每天来访游客的数量如下：id date visits 1 2017-07-01 100 …… 非常巧，id 字段刚好等于日期里面的几号。现在请筛选出连续三天都有大于 100 天的日期。 上面例子的输出为： date 2017-07-01 …… [SQL]**
    解析：
    select t1.date

```
from Tourists as t1, Tourists as t2, Tourists as t3
where (t1.id = (t2.id+1) and t2.id = (t3.id+1)) and t1.visits >100 and
t2.visits>100 and t3.visits>100
```