科大讯飞 1 组研究计划

一、 定义广告效果

经过讨论，我们初步决定分别用两种方式的指标来衡量广告效果，分别为：

1. **ROI**＝（销售额-广告投入成本）/广告投入成本*100%

   ROI＝ 总订单金额/总费用

   第二种计算方式为按数据字段说明中的计算方式。

   ROI 是最接近最后一步的指标，是最为重要的指标。

2. **综合各级转化率：**由于广告转化效果是漏斗式的，我们能收集到各层的转化率，如何从中选取就是一个值得思考的问题，对此，我们提出两种处理方式：

(1) **加权平均处理：**给出各个转化率权重，求其加权平均，将各转化率综合在一起。（权重来源：PCA，Lasso 等）

(2) **选择最差转化率：**用漏斗下层的人数除以上一层的人数，从中选取明显较低的一个座位广告效果。这样选择出的广告效果指标是最保守的，利用了短板效应的思路，同时还能看出用户购买或不购买录音笔的潜在原因是什么。(漏斗结果：广告投放量→点进落地页的人数→加入购物车/收藏/转发/停留超过一定时间的人数→购买人数）（停留时间要进行合理化数据处理）

3. 难点：

(1) 在进行加权时，各指标的权重如何分配可能需参考客户意见

(2) 除此之外是否还有其他客户认为重要的指标

二、 定义产品生命周期

观察产品销售波动情况，确定产品的实际生命周期，与公司预测的生命周期进行对比，优化公司新产品发布时间与产品迭代速度。具体实现途径如下：

1. **根据销售数据**中各产品的 2020-2021 年度全年销售情况，判断产品自投入市场以来的市场占有情况，以此**定义产品实际生命周期**。过程中需考虑引入购物节的平均销售量增长率来中和大型购物节（例如双十一，618）产生的影响，引入广告投放策略、广告投放总费用、广告投放效果等因素分析站内投放力度的变化对销售情况造成的影响。

2. 除此之外，由于新品的发布或错误的判断，可能导致过低地预测产品生命周期，过早下架产品。可以利用**时间序列模型**预测已下架产品未来可能的销售情况，结合实际销售数据，分析继续售卖该产品是否还有不错的市场前景，判断现有产品迭代速度是否符合市场变化，优化产品生命周期的确定。

希望补充数据：所有产品的发布时间与最终下架时间、科大讯飞预计的产品生命周期数据

三、 定义自变量

1. 目的：对数量繁多的自变量进行筛选，选出自变量数目合理且预测效果准确的模型，平衡预测模型的 bias 和 variance，得到最优预测模型。过于复杂的模型难以指导公司进行广告投放的策略改进，通过自变量的选择，指导公司结合经营状况与销售策略调整各类广告投放费用分配权重。

2. 步骤：

(1) 将衡量广告效果的变量，如单位广告成本促成销售量 or 用户转化率，作为因变量，将其他变量作为自变量，如广告成本，广告投放平台，广告类型（站内：搜索/展示，站外：品牌/效果），将分类变量转化为数值型变量，形成数据集。

(2) 将数据集按 3:7 随机分为训练集和测试集，使用**支持向量机（SVM）**建模，用训练集训练模型并调参，得到表现最好的参数组合。

(3) 使用模型对测试集数据进行预测，查看预测准确度。(若准确度欠佳可尝试**随机森林**或 **XGboost**，模型训练步骤与 2 一致）

(4) 对变量进行**敏感性分析**，得出不同自变量对因变量影响程度百分比排序，即得到自变量重要性排序，并据此调整广告投放费用分配权重。

四、 模型

1. 数据预处理：对数据进行简单分类，处理（清洗数据，处理异常值），合并

2. 数据处理：梳理（并拆分）数据时间段/站内站外渠道/手机端 PC 端

3. **渠道间协同/替代效应**（定量 test 存在难点，需要进一步讨论）：

(1) **定性方法**（例如作图观察各渠道数据的趋势、峰值是否接近，观察各渠道数据的总和是否保持相对稳定等）

(2) **定量 test**（取平均值后，看有无时间点显著区别于平均值，再对应该时间点进行研究和纵向比较，或许能发现差别），看数据是否有存在单独/两种/三种广告的时间段，A/B test（应用方式未确定，可能需要观察数据是否存在某一时间段只有一种或两种广告，与其他时间段全部类型广告都有的情况进行对比）

4. 模型及软件使用:

(1) 软件：Python, R, Excel

(2) 尝试模型：Tree，逻辑回归（非线性回归，可将效果通过门槛分成 1-达到预期和 0-未达到预期），Lasso（筛选因变量），PCA（效果-权重），机器学习，SVM 支持向量机，分类问题等。

5. 预测：尝试使用**时间序列模型**（应用在产品生命周期，验证产品下架是否为合理的操作，产品若不下架还能产生多少价值），对 20 年数据建模来预测 21 年数据，并与实际数据进行对比，需考虑 COVID19 影响（补充：或许可以对产品上架后的一段时间

进行观测，结果体现是否需要改变广告组合或者哪种广告组合更好）。

6. 建模方式

(1) 对分渠道、平台、时间等数据子集**分别建模**。

(2) **交乘项**：在筛选因变量、逻辑回归在、Lasso 中放入交乘项列，因为影响广告效果的因素众多，交乘项是有意义的。

7. 模型建好后的比较分析；

(1) **投放平台**：京东 VS 天猫

a) 整体分析

b) 按广告类型逐个对比分析

(2) **广告渠道**：PC 端 VS 移动端

(3) **广告类型**：如搜索类 VS 展示类

五、 站外：（潘媚）

站外广告我们分为品牌类和效果类两类来看。对于品牌类广告，我们关注各平台 KOL 发布广告内容后的互动效果，以 CPM、CPC 以及 CPE 为主要评判标准。对于效果类广告，我们依旧用 ROI 为指标来衡量广告效果。具体研究方式如下：

1. 品牌类：

由于品牌类效果转化很难具体衡量，我们考虑通过 CPM、CPC 以及 CPE 对比各平台同等级 KOL 的平均值，找出各个平台中互动效果好的 KOL，总结他们的特性，包括：主要创作领域、粉丝画像、文案、创作方式和风格等，得到精细化筛选达人的具体指标。尝试利用不同维度的具体指标去匹配找出优质达人。

2. 效果类：

对于京准通内数据，ROI 为主要的衡量指标。我们将这部分结合站内京东的数据一起通过前面所述的方式定义自变量再利用模型分析广告效果。

六、 评论销售：（吴亦君）

1. 评论数据分析

对于数据集中的评论数据，对**评论的数目、类型（好评差评等）乃至情感**进行分析，生成评论关键字的**词云**，来判定大众对产品的印象以及使用体验等，并作为影响销量和最终 ROI 等评判标准的因变量进行研究。

如果能够直接获取京东、天猫平台不同时间段的好评、差评、中评数据集，或许会对分析有一定帮助。

希望补充数据：评论数据的评论时间

2. 销售及促销数据分析

对于销售数据进行分层分析：

(1) 各渠道销售数据的分析与对比：

a) 通过**分析各渠道广告投放占比与最终销量占比间的关系**，对京东等六类渠道的销售数据分别进行分析，观测产品周期内不同时段，各渠道销量的变化趋势以及占比，探究这种变化与占比与广告投放的关系是否具有一定的普遍性；

b) 同时，格外关注新产品上线后，一定周期内不同渠道的销量变化，对比不同产品的情况，找出共性，区分差异，并从广告的投放力度、投放时长、不同平台的投放占比、投放投入成本等方面探究产生差异的原因，同时也要关注不同产品的目标人群、售价、功能等因素对销量的影响。

c) 关注不同 KOL、KOC 的宣传数据，尝试找出传播广、效率高的博主的特性。

(2) 对销售数据、广告数据等进行**逐月甚至逐周分析**：

以月报（或周报）的形式，展示销售数据随广告数据变化的情况，探究每个月销量亦或 ROI 等发生变化的原因，关注广告投放策略的变化，分析两者之间是否有一定关联。

通过对每月数据进行整理和分析，可以最终总结出广告投放与销量关联与否，并用于接下来的广告投放策略制定。但可能所需精力较多，有待考虑。

(3) 分析不同**促销力度的效果**：

观测促销期间对应的广告投放内容以及力度，探究促销力度以及促销活动的覆盖面（即宣传力度）对销量的影响，可以通过广告的曝光、点击等数值来判定覆盖面的大小，进行分析。

希望补充数据：销售目标量，用来和实际销售数据做对比

其他希望补充数据：

- ✓ **科大讯飞录音笔的用户画像；**
- ✓ **目前数据中京东包含 19 年数据而天猫、销售数据是从 20 年开始的，能否补充其他 19 年的数据？**
- ✓ **具体的各产品发布时间、下架时间（如有），产品型号间的主要区别，不同产品的主要目标人群、主要应用场景。**