

外呼通话营销对用户订购率的因果影响研究

——基于中国联通数据的实证分析

刘亮杰
香港科技大学

2025 年 5 月 16 日

引言

- 电信行业日益饱和且竞争激烈，提升用户响应率与订购率是运营商的重要目标。
- 外呼营销在实践中广泛应用，但其真实效果常受到用户画像、历史行为等混杂变量的干扰。
- 本研究提出基于因果推断的分析框架，结合倾向评分匹配与机器学习辅助的混杂变量识别方法。
- 引入双 XGBoost 模型结合 SHAP 解释，实现混杂变量的自动化识别；
- 将该方法应用于外呼策略，提供严谨的实证证据；
- 基于因果估计结果，提出可落地的精准营销策略建议。

相关研究

- 基于观测数据的因果推断在社会科学、医疗健康与营销等领域广泛应用。
- Rosenbaum 与 Rubin 提出的倾向评分匹配 (PSM) 方法成为控制混杂偏差的重要工具。
- 然而, 传统 PSM 方法使用逻辑回归对高维数据建模时, 难以捕捉复杂非线性关系与交互项, 且协变量选择易受主观影响。
- 近年来, XGBoost 等机器学习模型在捕捉变量关系、提升预测精度方面表现突出, SHAP 方法则提供了强可解释性的特征重要性分析。
- 本文提出融合 PSM、双机器学习模型与 SHAP 的因果推断框架, 并在电信营销中进行实证检验。

研究问题

本研究旨在回答以下核心问题：

- ① 外呼频率在多大程度上影响用户订购率？
- ② 通话时长是否显著影响订购行为？是否存在边际递减效应？
- ③ 控制混杂因素之后，外呼行为是否仍有显著因果影响？

为解答上述问题，研究设定以下目标：

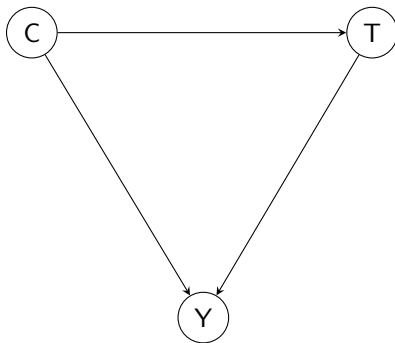
- 准确识别并控制影响因果估计的混杂变量；
- 构建基于倾向评分匹配（PSM）的因果推断模型；
- 定量评估外呼频率与通话时长对订购行为的处理效应；
- 基于因果效应分析结果，提出精准营销策略建议。

研究方法概述

- ① 基于有向无环图 (DAG) 明确因果结构;
- ② 使用双 XGBoost 模型结合 SHAP 值识别潜在混杂变量;
- ③ 构建倾向评分模型并进行 1:1 最近邻匹配;
- ④ 估计平均与异质性处理效应, 提出策略建议。

因果结构与识别假设

- 因果结构通过 DAG 图描述：用户特征 C 影响外呼策略 T 和订购结果 Y , T 直接影响 Y 。



图：因果结构 DAG 图

因果识别的关键假设

- SUTVA (稳定单元处理值假设): 个体的潜在结果不受他人处理状态影响;
- Unconfoundedness (条件独立性): 在控制协变量后, 处理与潜在结果独立;
- Overlap (公共支持): 每个个体接受与不接受处理的概率均非零。

混杂变量识别：双 XGBoost + SHAP 方法

- 为识别混杂变量，分别训练两个 XGBoost 模型：
 - **处理模型**：预测是否成功接通外呼；
 - **结果模型**：预测是否订购推荐套餐。
- 分别计算各变量的 SHAP 值（基于每个样本的特征边际贡献）；
- 在两个模型中均排名靠前的变量视为潜在混杂变量；
- 相比人工挑选变量，该方法更自动化、稳健。

为何使用 XGBoost + SHAP ?

- **XGBoost** 是梯度提升树 (Gradient Boosting Tree, GBT) 的高效实现, 优势包括:
 - 处理非线性关系与变量交互性强;
 - 自动处理缺失值与异常值;
 - 在高维结构中预测性能强于逻辑回归;
- **SHAP** 基于博弈论的 Shapley Value:
 - 能解释模型预测是由哪些特征推动的;
 - 保证一致性与局部准确性;
 - 对 XGBoost 等树模型具备专门优化算法。

倾向评分建模与匹配方法

- 倾向评分： $e(X) = \mathbb{P}(T = 1 | X)$ 表示接受处理的概率；
- 使用逻辑回归模型估计 $e(X)$ ；
- 匹配方法：1:1 最近邻匹配 + caliper = 0.2 (logit 标度)；
- 匹配后评估协变量平衡性，使用标准化均值差 (SMD) 判断：

$$\text{SMD} = \frac{\bar{X}_T - \bar{X}_C}{\sqrt{\frac{1}{2}(s_T^2 + s_C^2)}}$$

因果效应估计：ATE 与 ATT

- **ATE**（平均处理效应）：

$$ATE = \mathbb{E}[Y(1) - Y(0)]$$

- **ATT**（处理组平均处理效应）：

$$ATT = \mathbb{E}[Y(1) - Y(0) \mid T = 1]$$

- PSM 匹配样本主要代表处理组，因此研究重点放在 ATT。

Bootstrap: ATT 置信区间

- 为评估 ATT 的不确定性, 使用 Bootstrap 重抽样;
- 重复 $B = 1000$ 次抽样, 计算每次的 ATT 估计值:

$$\widehat{ATT}^{(b)} = \frac{1}{n} \sum_{i=1}^n \left(Y_i^{(T=1)} - Y_i^{(T=0)} \right)$$

- 用 2.5% 和 97.5% 分位数构建 95% 置信区间;
- 无分布假设, 适用于复杂匹配过程后的因果估计。

数据说明：基本信息

- 数据来源：中国联通 2024 年 12 月至 2025 年 1 月外呼营销项目；
- 样本规模：约 30 万名用户，87 万条外呼记录；
- 成功接通数约为 8 万次；
- 数据涵盖关键营销窗口，行为记录完整。

数据说明：变量分类

- **用户特征**：年龄、性别、客户类型、套餐类型、入网时长、历史行为等；
- **使用与支出行为**：语音、流量、短信总量，套餐外费用，是否超套；
- **处理变量**：当月外呼次数、接通次数、通话总时长；
- **结果变量**：是否订购推荐套餐（二值变量）、订购产品分类等。

未匹配样本分析：初步观察

- 评估外呼行为是否影响订购率，考察三类营销强度指标：
 - 外呼次数 (call_cnt)：外呼尝试频率；
 - 接通次数 (call_success_cnt)：成功触达用户的次数；
 - 通话时长 (call_hold_time)：累计通话持续时间（秒）。
- 初步观察发现：用户的中位数接通次数和通话时长均为 0，说明多数用户未被真正联系；
- 外呼资源消耗大，但实际“有效接触”比例较低；
- 接通与互动才是转化的关键，而不仅仅是拨打行为。

外呼行为的描述性统计

统计量	外呼次数	成功接通次数	接通率	通话时长（秒）
均值	2.84	0.26	23.48%	7.73
标准差	4.83	0.58	36.01%	40.58
中位数	1	0	0.00%	0.00
80 分位	6	1	50.00%	0.00
90 分位	7	1	100.00%	2.00
最大值	58	14	100.00%	2317

表：外呼行为统计特征（未匹配样本）

- 接通与通话行为高度偏态：80% 用户从未被接通，通话时长为 0；
- 呼叫尝试虽普遍存在，但互动非常有限。

订购率差异分析（未调整）

维度	有接触 (>0)	无接触 ($=0$)	差异
外呼次数	8.65%	2.23%	+6.42 个百分点
接通次数	17.40%	2.39%	+15.01 个百分点
通话时长	11.86%	4.77%	+7.09 个百分点

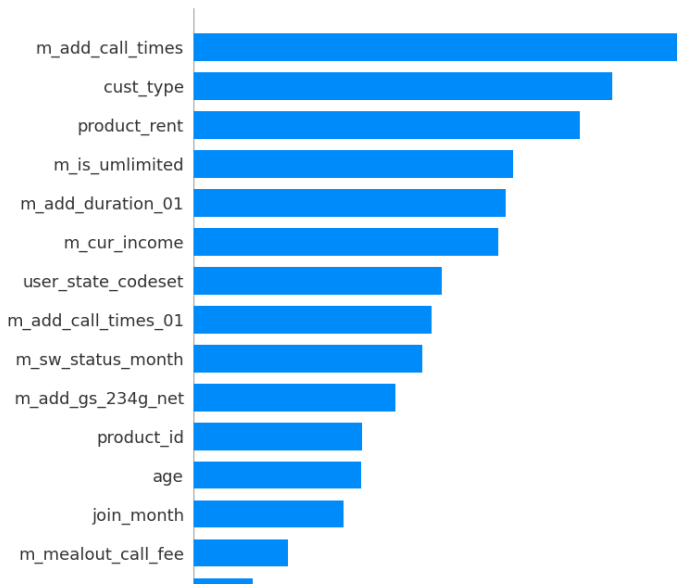
表: 不同接触强度下的订购率（未匹配样本）

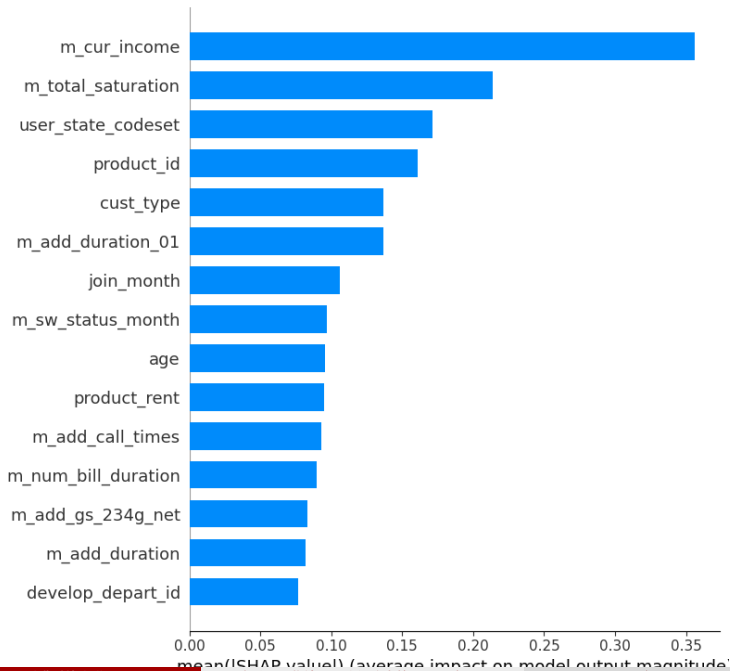
- 有效触达显著提升订购率，接通效果最强；
- 但此差异可能包含用户异质性偏差，尚不能断言因果关系；
- 后续需控制混杂变量并使用因果方法严谨估计。

混杂变量识别方法

- 训练两个 XGBoost 模型：
 - 处理模型：预测是否接通外呼；
 - 结果模型：预测是否订购产品。
- 计算每个模型中变量的 SHAP 值；
- 交集变量为潜在混杂变量，共识别出 12 项。

SHAP 排名前 15 特征 (示意图)





混杂变量示例（部分）

- age - 年龄影响外呼接通与订购；
- cust_type - 企业/个人客户策略不同；
- join_month - 入网时长影响处理概率与响应倾向；
- m_cur_income - 当前收入高的用户更可能被联系，也更可能订购；
- product_rent、product_id - 当前套餐影响外呼决策与订购决策。

匹配质量评估方法

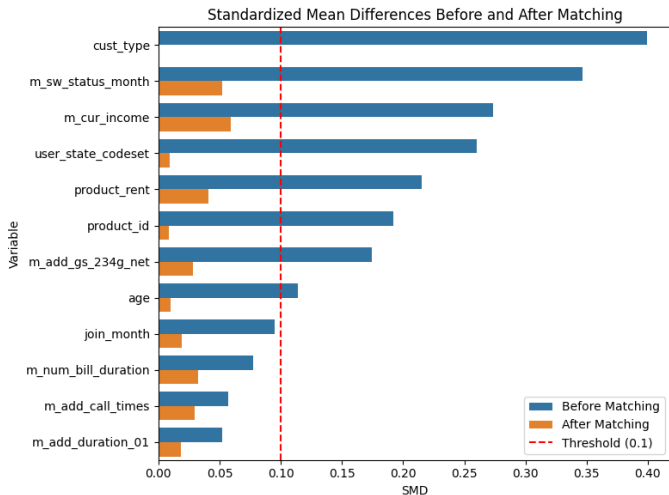
- 使用标准化均值差 (SMD) 衡量协变量匹配前后分布差异;
- 匹配前部分变量存在显著失衡 (如 `cust_type` 的 SMD 为 0.399);
- 匹配后所有协变量 SMD 均低于 0.1, 符合平衡标准。

匹配质量对比 (SMD 表格)

变量	匹配前 SMD	匹配后 SMD
cust_type	0.399	0.0009
m_cur_income	0.274	0.059
user_state_codeset	0.260	0.009
product_rent	0.215	0.041
age	0.114	0.010
join_month	0.095	0.019

表: 协变量匹配前后 SMD 对比

SMD 可视化图（建议）



图：匹配前后各协变量 SMD 对比图

因果效应估计方法：回顾

- 基于倾向评分匹配 (PSM) 后的样本, 估计处理效应;
- 聚焦三个外呼行为的二元处理定义:
 - ① 是否接到过外呼尝试 ($\text{call_cnt} > 0$);
 - ② 是否至少接通过一次外呼 ($\text{call_success_cnt} > 0$);
 - ③ 是否通话时长大于 0 ($\text{call_hold_time} > 0$)。
- 采用 Bootstrap 重抽样 ($B = 1000$) 构建 95% 置信区间;
- 重点分析 ATT (处理组平均因果效应), 具有实际营销指导意义。

ATT 估计结果与置信区间

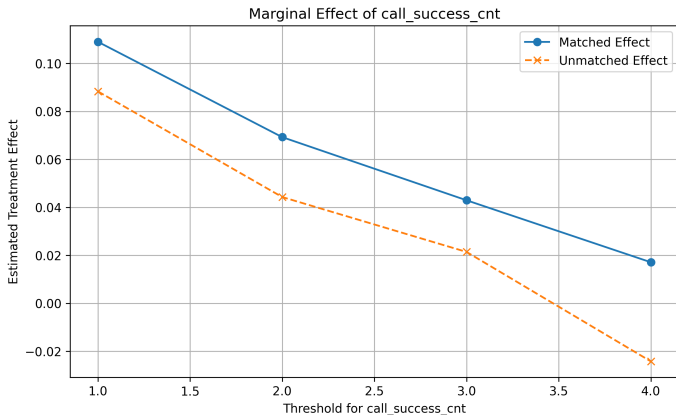
处理定义	匹配前差异	ATT	95% CI	偏差变化
外呼次数 (>0)	6.42%	6.64%	[6.49%, 6.79%]	+0.22pp
成功接通 (>0)	15.01%	15.24%	[14.91%, 15.56%]	+0.23pp
通话时长 (>0)	7.09%	8.97%	[8.58%, 9.38%]	+1.89pp

表: 三种处理定义下的 ATT 因果效应估计

- 所有外呼行为在控制混杂后, 均显著提升订购率;
- 成功接通效应最强, 通话时长效应次之;
- 匹配后 ATT 均高于未匹配差异, 表明原始策略未系统性偏好高响应人群。

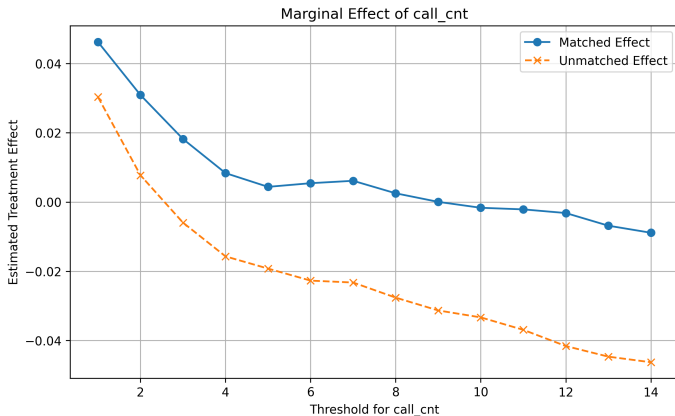
边际因果效应：接通次数

- 接通外呼次数与订购率呈非线性关系；
- 首次接通带来的边际效应最大，推动转化行为发生；
- 第 2-3 次仍有一定提升，但效应迅速递减；
- 多次重复拨打对同一用户作用有限，甚至可能引起反感。



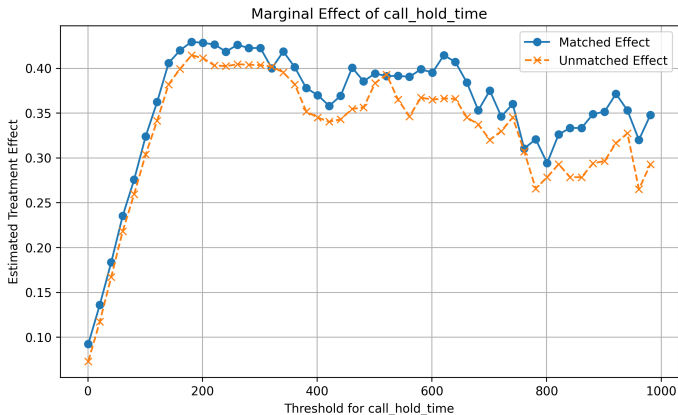
边际因果效应：外呼尝试频率

- 外呼频率在 1-3 次内具有明显正向效果；
- 超过 7-10 次后，**边际效应快速趋近于 0**，甚至出现负值；
- 说明过度外呼反而可能引起用户反感或疲劳；
- 强调“**覆盖面 > 过度触达**”的外呼策略导向。



边际因果效应：通话时长

- 较长通话（50–200 秒）显著提升转化概率；
- 超过 300 秒后边际效应波动，可能反映犹豫、抵触；
- “通得久”不等于“转化好”，应重视通话效率与质量；
- 沟通的关键在“打到点子上”，而非时间堆叠。



实证结论一：资源错配与潜力低估

- 所有处理定义下，匹配后的 ATT 均高于匹配前差异；
- 暗示当前外呼策略**未优先触达高响应个体**，存在资源投放偏误；
- 运营资源可能用于转化潜力低的“非响应人群”，而忽视了“被说服就订购”的高 uplift 用户；
- 建议基于**响应度预测（Uplift Modeling）**，优化外呼优先级排序机制。

实证结论二：边际分析的战略启示

- **接通优先**：一次成功接通最具价值，重复拨打收益下降；
- **频次控制**：外呼建议限制在 2-4 次，防止反向干扰；
- **优先接通**：成功接通转化率远高于尝试外呼；
- **时长把控**：通话控制在 50-200 秒区间为宜；
- **精准投放**：识别“高因果响应人群”，实现外呼 ROI 最大化；
- **建议构建基于 uplift 的用户分层模型**，提升整体营销投资效率。

研究局限

- **区域局限性**：数据来自中国联通单一区域，可能存在外推限制；
- **未观测混杂**：尽管控制了多个变量，仍可能存在未被记录的干扰因素；
- **PSM 局限性**：
 - ATT 估计代表处理组，无法代表整体；
 - 匹配过程中丢弃部分样本；
- **处理变量单一**：仅考虑了外呼频率与通话时长，未来可拓展至更多营销干预变量。

未来研究方向

- 引入更广泛的数据来源，覆盖全国用户样本；
- 探索工具变量 (IV)、双重机器学习 (DML) 等稳健因果识别方法；
- 结合逆概率加权 (IPW)、TMLE 等估计策略；
- 应用强化学习框架，设计动态外呼策略；
- 拓展至其他营销场景，如套餐升级、留存激励等。

感谢聆听！

欢迎提问

刘亮杰
香港科技大学