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 15 日

摘要

在电信行业高度饱和且竞争激烈的市场环境中，提升用户订购率成为关键任务。外呼营销作为直接触达用户的重要渠道，其真实的因果效应却常常受到混杂因素影响而难以准确识别。本研究结合倾向评分匹配（PSM）与基于双 XGBoost 模型与 SHAP 解释的混杂变量识别方法，实证分析外呼频率与通话时长对用户订购行为的因果影响。结果显示，较高的外呼频率与更长的通话时长均可显著提升订购率，且呈现边际递减与最佳通话时长窗口。本研究为电信行业精准外呼策略制定提供了实证支撑与方法借鉴。

关键词：因果推断、外呼营销、订购率、倾向评分匹配、SHAP、机器学习

1 引言

在电信行业日益饱和且竞争激烈的背景下，提升用户响应率与订购率是运营商的重要目标。作为直接触达用户的手段，外呼营销在实践中广泛应用。然而，其真实效果常受到用户画像、历史行为等混杂变量的干扰，单纯统计分析难以揭示其因果机制。

本研究提出一套基于因果推断的分析框架，结合倾向评分匹配（PSM）与机器学习辅助的混杂变量识别方法，系统评估外呼频率与通话时长对用户订购率的因果影响。

具体贡献如下：首先，创新性地引入双 XGBoost 模型结合 SHAP 解释，实现混杂变量的自动化识别，增强倾向评分建模的可信度；其次，将该方法应用于电信行业外呼策略这一实践中重要但研究稀缺的课题，提供严谨的实证证据；最后，基于因果估计结果，提出可落地的营销策略建议，为精准外呼提供理论依据与实践指导。

2 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得到了广泛研究。Rosenbaum 与 Rubin 提出的倾向评分匹配（PSM）方法已成为控制混杂偏差的重要工具，通过在处理组与控制组之间平衡协变量来估计因果效应 [1]。

然而，传统 PSM 方法在处理高维数据时面临局限。基于逻辑回归的倾向评分模型往往难以捕捉复杂的非线性关系与交互项，可能存在残余混杂 [2]；同时，协变量选择依赖专家经验，容易引入主观偏差 [3]。

近年来，机器学习方法被逐步引入因果推断框架。XGBoost 等梯度提升模型在处理变量间复杂关系与提升预测精度方面表现优异 [4]；而 SHAP（SHapley Additive Explanations）方法则提供了可解释性强的特征重要性分析工具 [5]。

尽管已有研究尝试将机器学习方法用于倾向评分建模，但系统性地结合双模型与 SHAP 解释以辅助混杂变量识别的实证研究仍较为缺乏。为此，本文提出一套融合 PSM、双机器学习模型与 SHAP 变量选择的统一因果推断框架，并在电信外呼营销应用场景中进行了实证验证。

3 研究问题与研究目标

3.1 研究问题

本研究主要探讨以下问题：

- 外呼频率在多大程度上影响用户订购率？
- 外呼通话时长是否影响订购行为？是否存在边际递减效应？
- 在控制混杂因素之后，外呼行为是否仍对用户订购决策存在显著因果影响？

3.2 研究目标

为解答上述问题，研究目标包括：

- 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的处理效应，并分析其商业意义；
- 通过严谨的实证检验评估估计结果的稳健性与可推广性。

4 研究方法

本研究方法包括以下四个步骤：首先，基于有向无环图（DAG）明确因果结构；其次，通过双 XGBoost 模型结合 SHAP 值识别潜在混杂变量；第三，构建倾向评分并进行 1:1 最近邻匹配；最后，估计平均与异质性处理效应，提出策略建议。

4.1 因果结构与识别假设

本研究因果框架基于潜在结果模型，并通过 DAG 图表示变量之间的假设因果关系。如图 1 所示，用户画像与历史行为（**C**）影响外呼策略（**T**）与订购结果（**Y**），而外呼策略直接影响订购结果。

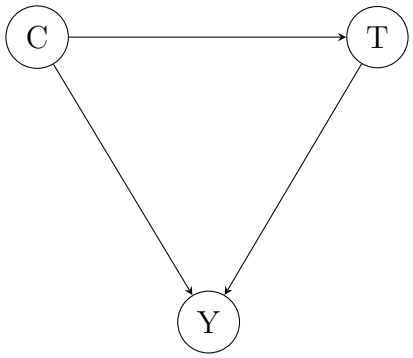

图 1: 因果结构的 DAG 图表示

为确保基于观测数据进行因果推断的合理性，本研究遵循以下假设：

- **稳定单元处理值假设（SUTVA）**：一个个体的处理状态不影响其他个体的潜在结果；
- **条件独立性假设（Unconfoundedness）**：在控制协变量后，处理分配与潜在结果独立；
- **公共支持假设（Overlap）**：每个个体都有接受与不接受处理的正概率。

以上假设为使用倾向评分进行因果效应估计提供了理论基础。

4.2 基于双机器学习模型与 SHAP 值的混杂变量识别

针对用户行为数据维度高、变量间关系复杂的特征，本研究使用一种结合双机器学习模型与 SHAP（SHapley Additive Explanations）值的混杂变量识别方法。我们分别训练两个 XGBoost 模型：一个用于预测处理分配（如接受高频外呼的概率），另一个用于预测订购结果（即是否订购推荐套餐）。

XGBoost 是梯度提升树（Gradient Boosting Tree, GBT）的高效实现，具备建模非线性关系、变量交互效应与缺失值鲁棒处理能力，在实际预测中表现优异。然而，作为集成模型，XGBoost 存在“黑箱”问题，不易直观解释模型中各特征对预测的具体贡献。

为此，我们引入 SHAP 方法对模型结果进行解释。SHAP 值基于博弈论中的 Shapley 分配原则，能够将模型预测合理地分配到每个输入特征，具有一致性与局部准确性，特别适用于树模型的解释。

在处理模型与结果模型中，分别计算每个特征在所有样本上的 SHAP 值，并进行重要性排序。若某变量在两个模型中均具有较高排名，则视为潜在混杂变量，因其可能同时影响外呼分配与用户订购决策。

该方法结合了机器学习的自动化选择能力与 SHAP 的可解释性，相较传统依赖主观经验的变量筛选方式，更加透明、稳健，有助于在高维观测研究中提升因果效应估计的有效性。

4.3 倾向评分建模与因果效应估计

4.3.1 倾向评分建模与匹配方法

倾向评分（Propensity Score）记为 $e(X) = \mathbb{P}(T = 1 \mid X)$，表示在给定协变量 X 的条件下个体接受处理（$T = 1$）的概率。本文采用逻辑回归模型对倾向评分进行估计，其中处理变量（如接受高频外呼）作为因变量，混杂变量作为解释变量。

倾向评分估计完成后，我们采用 1:1 最近邻匹配法（Nearest Neighbor Matching），并引入 caliper 限制匹配质量。具体而言，每个处理组个体与倾向评分最接近的对照组个体进行匹配，且两者倾向评分差距不得超过其 logit 转换值标准差的 0.2 倍（即 caliper = 0.2），该标准符合主流实证研究建议。此方法有助于避免低质量匹配，减少估计偏差。

为评估匹配效果，采用标准化均值差（Standardized Mean Difference, SMD）作为协变量平衡性指标，其定义如下：

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

其中 $\bar{X}_T$ 与 $\bar{X}_C$ 分别为处理组与对照组协变量的样本均值，s_T^2 与 s_C^2 为相应样本方差。

本研究设定 SMD 阈值为 0.1，匹配后所有协变量的 SMD 若小于此值，则视为协变量分布已充分平衡，可进行后续因果效应估计。

4.3.2 因果效应估计

在完成倾向评分匹配后，我们基于潜在结果模型（Potential Outcomes Framework）对因果效应进行了估计。本研究重点关注以下两个估计量：

- **ATE（总体平均处理效应）**：

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

表示如果总体中所有个体都接受处理与都未接受处理之间的平均潜在结果差异。

- **ATT（处理组平均处理效应）**：

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

表示在实际接受处理的个体中，平均因果效应的大小。

由于本研究采用倾向评分匹配（PSM）以处理组为基础构建匹配样本，因此匹配后的样本代表的是处理组子样本，而非整体总体。因此，我们主要关注 ATT 的估计，因其在匹配样本中可识别性较强，解释也更具实用价值。相比之下，ATE 在处理组与未处理组之间的支持区域重叠有限时，可能出现偏差。

为评估 ATT 估计的不确定性，我们引入了非参数 Bootstrap 重抽样方法。设匹配后的样本中包含 n 对处理组与对照组的匹配对，其估计流程如下：

1. 对于每一次 Bootstrap 迭代 $b = 1, \ldots, B$（本研究设 $B = 1000$），我们从原始匹配样本中以有放回的方式抽取 n 对匹配对，构成一个重抽样样本。

2. 对每个重抽样样本，计算 ATT 的估计值：

$$\widehat{\mathrm{ATT}}^{(b)} = \frac{1}{n} \sum_{i=1}^{n} \left(Y_i^{(\text{处理组})} - Y_i^{(\text{对照组})} \right),$$

其中 $Y_i^{(\text{处理组})}$ 和 $Y_i^{(\text{对照组})}$ 分别表示每对样本中的处理组与对照组观察到的订购结果。

3. 完成 B 次迭代后，得到估计值集合 $\{\widehat{\mathrm{ATT}}^{(1)}, \ldots, \widehat{\mathrm{ATT}}^{(B)}\}$，以其经验分布的第 2.5% 与 97.5% 分位数构造 95% 置信区间。

该 Bootstrap 方法不依赖于特定分布假设，能够较为稳健地反映因匹配与抽样过程带来的估计不确定性。

除平均处理效应外，本研究还进一步探讨了边际处理效应（Marginal Treatment Effect）与异质性处理效应（Heterogeneous Treatment Effect），以识别用户对外呼频率与通话时长的非线性响应模式及不同子群体间的反应差异。这些结果对制定更具针对性的外呼营销策略具有重要实践价值。

5 数据说明

研究所用数据来自中国联通 2024 年 12 月至 2025 年 1 月的外呼营销项目，包含用户画像、历史行为、外呼活动及订购情况。

- **样本规模**：约 30 万名用户，含约 87 万条外呼记录，其中约 8 万次为成功接通；
- **观察时间段**：2024 年 12 月至 2025 年 1 月，覆盖关键营销窗口，确保行为充分观测；
- **变量概览**：
  - **用户特征**：年龄、性别、客户类型、套餐类型、入网时长、历史充值与停机行为等；
  - **使用与支出行为**：语音、流量、短信使用总量、套餐外费用、是否超套等指标；
  - **处理变量**：当月外呼次数与接通次数，用以衡量干预强度；
  - **结果变量**：是否成功订购推荐套餐（二值变量），并记录所订套餐类型与产品分类。
- **隐私合规性**：所有数据在处理前已脱敏处理，符合用户隐私与数据合规相关规定。

6 实证结果

6.1 未调整偏差分析

为了评估是否有必要控制混杂因素，我们首先对全体未匹配样本进行分析，从三个维度衡量外呼营销强度：**外呼频率**（即外呼次数 `call_cnt`）、**接通次数**（成功接通的外呼数量 `call_success_cnt`）、以及**通话时长**（外呼累计通话时长 `call_hold_time`，单位秒）。这三项指标分别反映了从尝试接触、成功联系，到持续互动的不同层级的营销投入。

表 1 总结了这些外呼行为变量的分布情况。平均来看，用户在观测期内接到 2.84 次外呼，但接通次数和通话时长的中位数均为 0，说明超过一半的用户从未被成功联系。偏态特征尤为显著地体现在接通维度上：即便在 80 分位数，用户的成功接通次数也仅为一次，通话时长仍为 0。直到 90 分位数，通话时长才首次出现显著值（大于 2 秒）。这表明虽然呼出数量不低，但真正实现的用户互动极为有限。

为了分析是否“有”接触即能带来更高订购率，我们为每个维度构造了二元处理变量，值为 1 表示该用户在该维度上数值大于 0，0 表示未有任何接触。表 2 比较了不同处理状态下的订购率。

结果显示显著差异：曾收到过至少一次外呼的用户，其订购率为 8.65%，远高于从未被联系过用户的 2.23%。如果进一步考虑联系质量，差异更为明显：至少接通一次的用户，其订购率高达 17.40%，而从未接通的用户仅为 2.39%。同样，在通话时长大于 0 的用户中，订购率为 11.86%，而未有通话的用户仅为 4.77%。

这些现象凸显了“有效接触”的重要性，而不仅仅是尝试呼叫本身。然而，这些未经调整的差异也可能受到选择偏差的影响：更容易接触到或被优先外呼的用户，可能本身在人群特征或历史行为上就有所不同。因此，这些结果具有提示意义，但并不能直接支持因果解释，仍需后续使用匹配方法控制混杂变量后进行严谨的效应估计。

表 1: 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
60 分位	2	0	8.00%	0.00
70 分位	4	0	25.00%	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 2: 是否有接触与订购率对比（未匹配样本）

维度	处理组（>0）	控制组（=0）	差异
外呼次数	8.65%	2.23%	+6.42 个百分点
成功接通次数	17.40%	2.39%	+15.01 个百分点
通话时长	11.86%	4.77%	+7.09 个百分点

6.2 识别混杂变量

为了识别那些可能同时影响用户是否被成功外呼和是否转化订购的协变量，我们采用了双模型 SHAP（SHapley Additive exPlanations）解释框架。具体而言，我们分别训练了两个 XGBoost 分类模型：

- 一个**处理模型**用于预测用户是否接到至少一次成功外呼（即 `call_success_cnt > 0`）；
- 一个**结果模型**用于预测用户是否办理了目标产品（即 `is_order = 1`）。

我们计算了每个模型中协变量的 SHAP 值，用以衡量其边际贡献。然后按照 SHAP 绝对值的均值对变量重要性进行排序，并分别保留每个模型排名前 15 的变量。两个集合的交集即为**高优先级混杂变量**：即同时高度预测处理与结果的变量，共识别出 12 项。

这些变量涵盖了用户的行为、人口统计特征以及套餐信息，具体包括：

- **age** –年龄较大的用户可能更忠诚、关注度更高，但也可能更难通过电话联系。
- **m_add_duration_01** –主叫通话总时长越高，越可能被认定为活跃用户，但也可能反映可接触性。
- **join_month** –入网时长越久的用户更可能成为营销重点，同时其转化倾向也更高。
- **m_num_bill_duration** –累计计费语音时长可能反映套餐匹配度和通信需求。
- **cust_type** –企业用户通常更易被联系（如有固定联系人），也更倾向采纳业务。
- **product_id** –所属套餐反映用户细分、历史优惠或过往营销策略。
- **m_add_gs_234g_net** –用户使用的最高网络类型（如 5G）同时影响外呼对象选择与产品适配度。
- **m_sw_status_month** –近月 5G 开关使用情况可能表征 churn 风险或兴趣，影响外呼及订购。
- **m_add_call_times** –通话活跃度越高可能意味着价值高、响应率高。
- **product_rent** –高资费套餐用户往往更可能获得营销资源，同时也更有购买意愿。
- **m_cur_income** –收入高的用户是外呼优先人群，其自然转化倾向也更强。
- **user_state_codeset** –用户当前状态（如是否停机）影响其是否可被成功联系。

这些变量体现了电信运营商在外呼策略中所依赖的用户分群逻辑。例如高消费用户（`m_cur_income`）或老用户（`join_month`）可能因其高价值被优先联系，同时其订购倾向也较高。如果在因果估计中忽视这些变量，可能会误将其自然行为归因于外呼的“效果”，从而造成偏误。

因此，在倾向得分模型中纳入这些变量，对于减少混杂偏差、确保处理组和控制组的可比性是必不可少的。

表 3 展示了两个模型中排名前 15 的 SHAP 特征。图 2 和图 3 分别展示了它们在预测处理和结果中的重要性排序图。

表 3: 处理模型与结果模型的前 15 个 SHAP 特征

处理模型变量	**SHAP 值**	**结果模型变量**	**SHAP 值**
m_add_call_times	0.239	m_cur_income	0.356
cust_type	0.205	m_total_saturation	0.214
product_rent	0.189	user_state_codeset	0.171
m_is_umlimited	0.157	product_id	0.161
m_add_duration_01	0.153	cust_type	0.137
m_cur_income	0.149	m_add_duration_01	0.137
user_state_codeset	0.122	join_month	0.106
m_add_call_times_01	0.117	m_sw_status_month	0.097
m_sw_status_month	0.112	age	0.096
m_add_gs_234g_net	0.099	product_rent	0.095
product_id	0.082	m_add_call_times	0.093
age	0.082	m_num_bill_duration	0.090
join_month	0.073	m_add_gs_234g_net	0.083
m_mealout_call_fee	0.046	m_add_duration	0.082
m_num_bill_duration	0.029	develop_depart_id	0.077

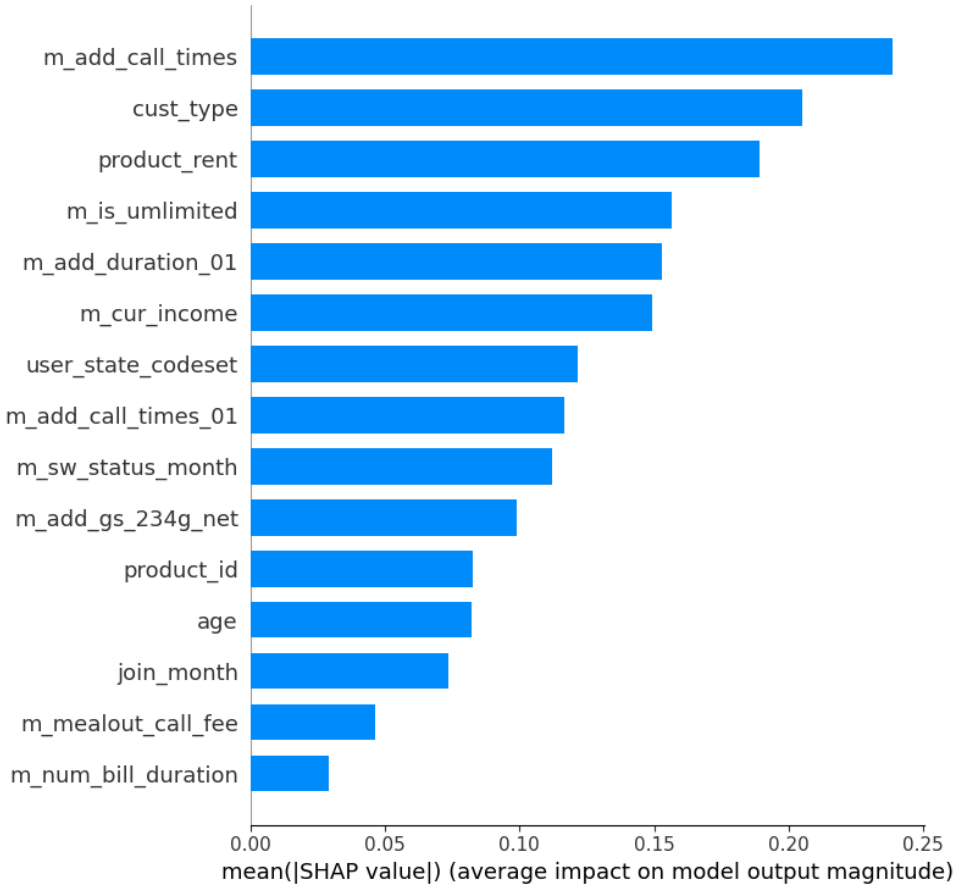

图 2: SHAP 解释图–处理模型（前 15 特征）

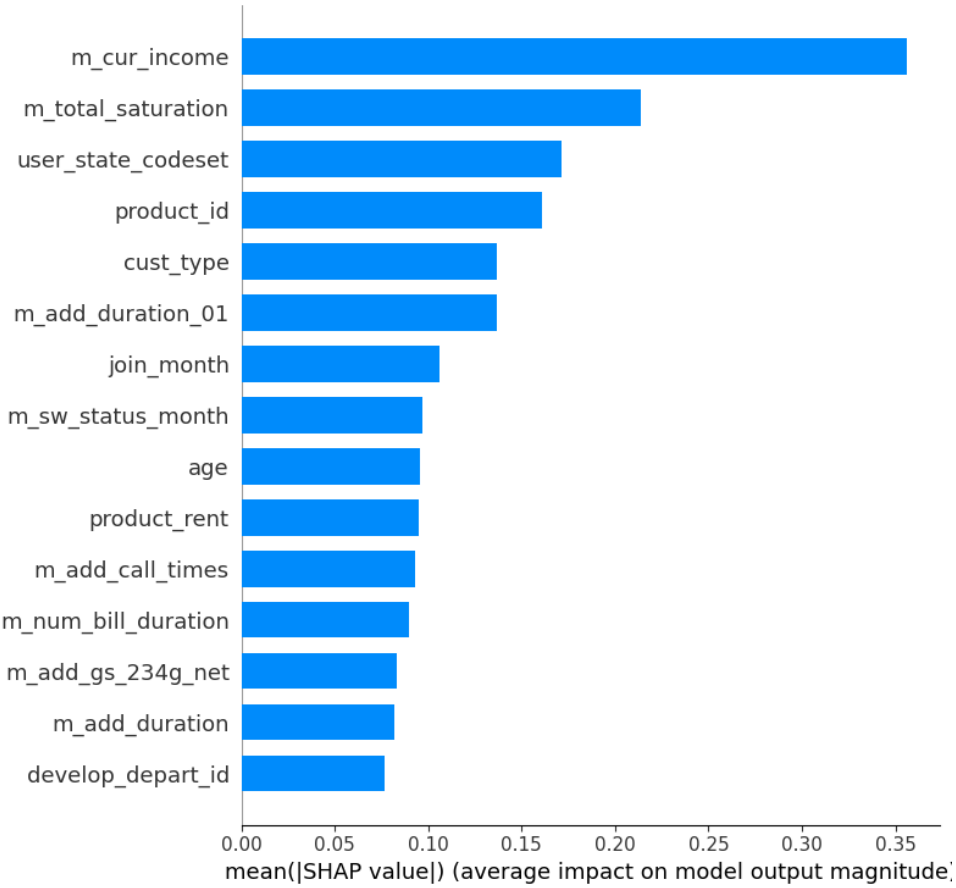

图 3: SHAP 解释图–结果模型（前 15 特征）

6.3 匹配质量评估

为了解决潜在的混杂偏差，我们采用了倾向得分匹配（Propensity Score Matching, PSM），所使用的 12 个协变量由前述双模型 SHAP 分析得出。这些变量涵盖了用户人口属性、套餐信息和行为特征，例如 `cust_type`、`product_rent`、`m_cur_income` 等。目标是构建在关键协变量上具有统计可比性的处理组与控制组。

我们使用**标准化均值差异（Standardized Mean Differences, SMD）**来评估匹配效果。如图 4 所示，匹配前后每个协变量的 SMD 值对比清晰地揭示了协变量平衡性的提升。我们采用了 SMD < 0.1 作为可接受平衡的判断标准。

在匹配之前，部分变量存在显著不平衡：

- `cust_type`（SMD $= 0.399$），说明用户类型之间存在显著分群；
- `m_sw_status_month`（SMD $= 0.346$），体现 5G 使用行为在组间分布不均；
- `m_cur_income`（SMD $= 0.274$），可能反映收入差异引发的外呼优先级偏差。

而在匹配之后，所有变量的 SMD 均降至 0.1 以下，大多数甚至低于 0.05，说明处理组与控制组在协变量分布上已高度可比。例如，`cust_type` 的不平衡从 0.399 降至

0.0009，`m_cur_income` 从 0.274 降至 0.059。

这种改进充分表明了我们匹配策略的有效性，并为后续的因果效应估计提供了稳固的基础。

表 4: 匹配前后标准化均值差异（SMD）比较

协变量	匹配前 SMD	匹配后 SMD
cust_type	0.399	0.0009
m_sw_status_month	0.346	0.0524
m_cur_income	0.274	0.0594
user_state_codeset	0.260	0.0092
product_rent	0.215	0.0406
product_id	0.192	0.0089
m_add_gs_234g_net	0.174	0.0279
age	0.114	0.0097
join_month	0.095	0.0189
m_num_bill_duration	0.078	0.0326
m_add_call_times	0.057	0.0295
m_add_duration_01	0.052	0.0183

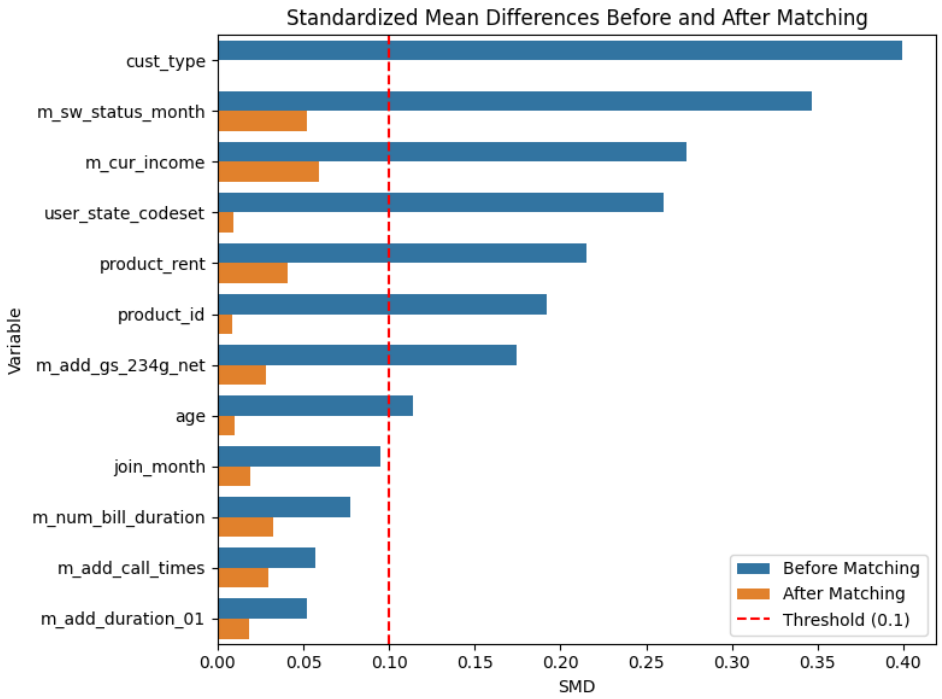

图 4: 匹配前后各协变量的标准化均值差异（SMD）

6.4 因果效应估计

为了量化外呼行为对用户订购行为的因果影响，我们开展了两部分分析。第一步，我们使用倾向得分匹配（Propensity Score Matching, PSM）估计了三种二元处理定义下的**平均处理效应（ATT）**。第二步，我们通过设定不同的阈值，考察处理强度变化下的边际因果效应，以揭示潜在的非线性响应模式。

一、二元处理效应分析（Binary Treatment Effects） 本研究设定了三种二元处理变量，分别为：**（1）是否曾接到任何外呼**（`call_cnt > 0`）；**（2）是否至少接通一通外呼电话**（`call_success_cnt > 0`）；**（3）是否累计通话时长大于 0 秒**（`call_hold_time > 0`）。表 5 汇总了各处理定义下的倾向评分匹配结果。

从匹配后的平均处理效应（ATT）估计来看，三种外呼行为均显著提升了用户的订购概率。尤其是“成功接通外呼”这一处理变量，其 ATT 为 15.24%，且 Bootstrap 置信区间为 [14.91%, 15.56%]，显著高于其他两项处理的因果效应。相比之下，“仅有外呼尝试”对应的 ATT 为 6.64%（CI: [6.49%, 6.79%]），“通话时长大于 0”对应的 ATT 为 8.97%（CI: [8.58%, 9.38%]）。这进一步表明，**达成有效接触**远比单纯增加**外呼次数**更具实质性营销价值。

表 5: 倾向得分匹配后的因果效应估计（含置信区间）

处理类型	匹配前差异	ATT（匹配后差异）	95% 置信区间	偏差减少
外呼频率（`call_cnt > 0`）	6.42%	6.64%	(6.49%, 6.79%)	−0.22 pp
外呼接通（`call_success_cnt > 0`）	15.01%	15.24%	(14.91%, 15.56%)	−0.22 pp
通话时长（`call_hold_time > 0`）	7.09%	8.97%	(8.58%, 9.38%)	−1.89 pp

二、边际因果效应分析（Marginal / Dose-Response Effects） 为了进一步探索外呼强度对订购效果的影响，我们依次提高处理定义的阈值，并分别估计匹配前与匹配后的订购率差异，从而绘制剂量-响应曲线。图 5、6 和 7 分别展示了外呼接通次数、外呼尝试次数与通话时长这三个维度下的边际因果效应。

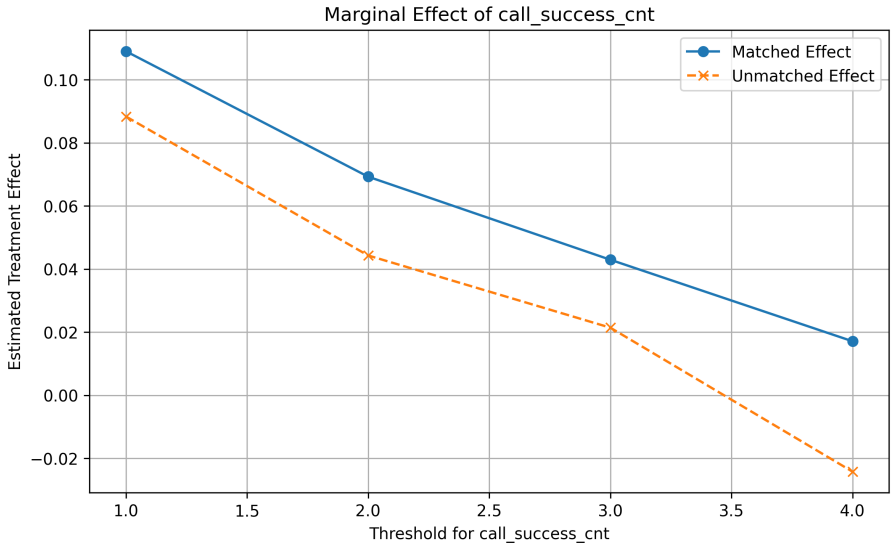

图 5: 不同接通次数下的边际处理效应

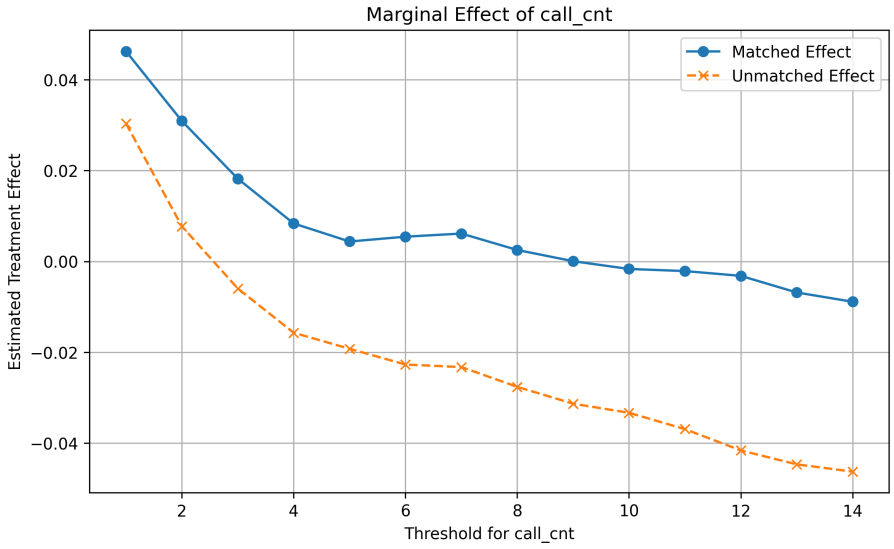

图 6: 不同外呼次数下的边际处理效应

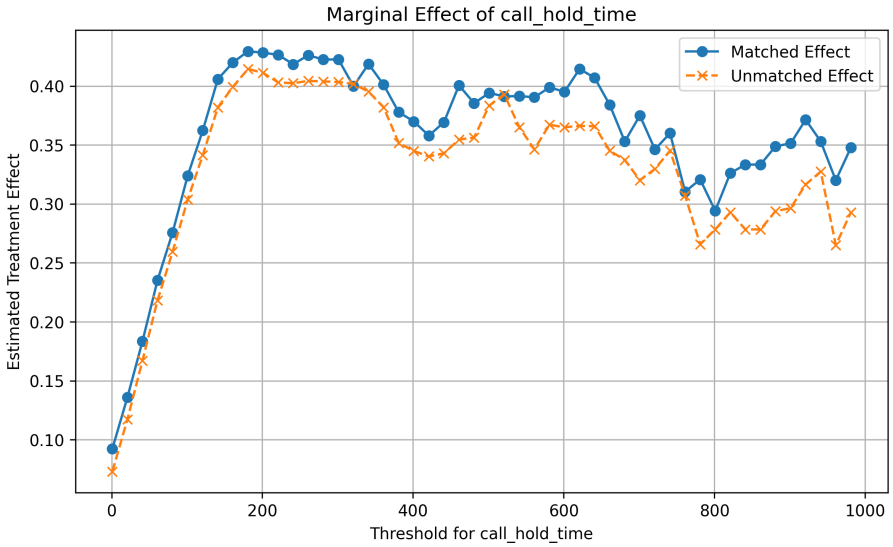

图 7: 不同通话时长下的边际处理效应

6.5 实证结论

一、**外呼策略存在资源错配风险** 上述分析揭示了一个关键的运营洞察：在所有三种处理定义下，**匹配后的平均处理效应（ATT）均高于未匹配样本中的订购率差异**，对应的偏差方向均为负。这一现象说明，当前的外呼策略**并未系统性地优先覆盖响应意愿更强的用户**，甚至可能存在营销资源错配的问题。

进一步解读，该结果表明：原始外呼分配机制可能依据订购率或基础画像进行策略制定，**而非基于外呼带来订购提升的“响应度”进行精准干预**。换言之，营销资源可能被投入到了订购率本就较低、且难以通过外呼改变行为的“非响应人群”，而非那些真正“被说服即可订购”的目标人群（即具有高因果响应效应者）。

因此，我们建议未来外呼策略设计应引入**响应评分模型**（uplift modeling），识别那些在接受外呼后最有可能产生转化的用户（即 high-uplift 个体）。同时结合用户的活跃度、收入水平与历史行为等特征，构建**以因果提升为核心的外呼优先级排序机制**，以最大化整体营销投资回报（ROI），并避免资源浪费与用户干扰带来的负面影响。

二、**来自边际效益分析的战略启示** 剂量-响应图（Figures 5–7）揭示了多个与营销策略密切相关的重要规律：

- **外呼接通次数：** 当用户首次成功接通外呼时，订购率提升最为显著；而在第 2 3 次之后，边际效应开始递减。这表明一次有效沟通往往足以促成转化，重复拨打对同一用户的边际价值有限。
- **外呼尝试次数：** 订购效应在 1 3 次呼叫范围内有所上升，但在超过 7 10 次之后迅速下滑，甚至趋于负值。这暗示过度外呼可能引发用户反感，特别是当目标人群属于低响应或非目标群体时。

- **通话时长：**较长的通话时间（大约 200 秒以内）与较高的订购概率正相关。在该阈值之后，效应趋于平缓甚至波动，提示**沟通质量**比**通话时间长度**本身更为重要。超长通话可能反映用户犹豫、不确定，甚至反感。

上述规律强调了外呼营销的**非线性本质**。要避免资源浪费和用户体验下降，营销人员应在**频次**与**质量**之间精准平衡，确保每一通电话都具备实质性价值。

7 实践建议

基于前述因果推断结果，我们提出以下可操作的外呼优化策略：

- **优化外呼频率：**每位用户的外呼次数应控制在 2–4 次之间。超过该范围后，转化提升效应显著下降，甚至可能因骚扰体验引起用户反感或投诉。
- **关注首次接通：**确保每位用户至少有一次有效接通。大部分转化效果来自首次触达，说明**覆盖面**比**重复投放**更关键。
- **控制通话时长：** 50–200 秒之间的通话最具转化潜力。过短可能沟通不足，过长则可能导致信息过载、用户疲劳或犹豫情绪。
- **优先高转化可能性的人群：**建议建立用户评分模型，**优先触达高转化可能性**、**价值高**、**响应意愿强的群体**，以实现资源最大化配置。

上述建议可为运营商制定数据驱动的外呼策略提供指导，兼顾营销效率与用户体验，提升订购转化效果并优化人力与资源成本。

8 研究局限与未来方向

尽管本研究取得了较为明确的实证结论，但仍存在以下局限，需在未来研究中予以重视和拓展：

首先，分析基于中国联通某一区域分公司的数据，样本的地理与业务环境具有一定局限性，可能限制了研究结论的外推性。未来可引入多区域、全国性样本，以验证结果的稳健性与普适性。

其次，尽管本研究通过倾向评分匹配和机器学习变量选择等手段尽可能控制了混杂因素，但仍无法完全排除潜在的未观测混杂变量干扰。未来建议结合工具变量（IV）、双重机器学习（DML）等方法，进一步提升因果识别的稳健性。

第三，倾向评分匹配（PSM）在提升协变量平衡性方面表现良好，但其本质上更适用于估计处理组的平均处理效应（ATT），且匹配过程中可能丢弃部分样本，影响估计精度和代表性。此外，当倾向评分模型设定不当时，PSM 可能引入新偏差。未来研究

可考虑引入逆概率加权（IPW）、目标最大似然估计（TMLE）或双重稳健估计等方法，进一步强化因果推断结果。

第四，当前研究聚焦于基础订购场景中的外呼频率与通话时长两类变量。未来可将研究拓展至其他营销场景，如流量包升级、套餐捆绑推荐、用户留存激励等。同时，可结合强化学习、序贯决策等框架，探索动态适应式外呼策略在实际业务中的部署与优化路径。

9 结论与启示

本研究基于中国联通实证数据，识别了外呼通话频率与时长对用户订购率的显著因果效应。结果显示，适度的外呼频率（2–4 次）与中等通话时长（50–90 秒）能够显著提升用户转化率，且在边际层面呈现出递减或非线性特征。

上述结论为电信企业外呼策略制定提供了清晰的实务启示：在避免过度打扰的前提下，精准把控外呼强度与沟通节奏，是提升订购率、优化客户体验的关键所在。

在方法方面，本文构建了一个稳健的因果推断分析框架，结合倾向评分匹配与基于 XGBoost + SHAP 的变量选择机制，在应对高维协变量混杂时表现出良好稳定性与可解释性。该混合方法论可被推广应用至其他行业场景，如保险、零售、金融等领域的行为干预评估。

未来研究可在本研究基础上进一步探索跨情境验证、估计方法对比与动态外呼策略等方向，推动因果推断方法在商业决策中的深入应用与发展。

参考文献

[1] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[2] Guido W Imbens and Donald B Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015.

[3] Susan Athey and Guido W Imbens. The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, 31(2):3–32, 2017.

[4] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM, 2016.

[5] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, pages 4765–4774, 2017.