广东证券期货业协会 2024 年
重点课题研究报告

课题名称：　　因果推理助力零售业务营销的探索实践

课题研究单位:　广发证券股份有限公司

课题负责人:　　辛治运

课题组成员:　　王蓁，黄紫菲，张肖，彭文慧，黄敏莎，蔡珠清，缪金朋

课题研究时间：　2024 年 9 月-2024 年 12 月

目录

第一章　课题背景

第一节 因果推理信创方案简要介绍

在数字化转型的浪潮席卷全球的当下，信创产业蓬勃发展。信创产业的核心目标在于实现信息技术领域的自主可控和安全可靠，这不仅关乎国家的信息安全，更是推动各行业数字化转型的重要保障。

本研究提出的因果推理信创方案，将信创理念与前沿的因果推理技术深度融合，为证券行业智能运营策略的制定提供了新的解决方案。

因果推理，作为一门专注于研究如何确定变量之间因果关系的学科，在数据分析领域有着举足轻重的地位。在传统的数据分析模式中，我们往往更多地关注变量之间的相关性。然而，一个重要的认知是，相关性并不等同于因果关系。就像一个有趣的例子，冰淇淋销量和游泳人数之间呈现出正相关关系，即冰淇淋销量上升时，游泳人数也随之增加。但这并不意味着冰淇淋销量的增加会直接导致游泳人数的上升。实际上，它们可能都受到气温升高这一共同因素的影响。气温升高使得人们对冰淇淋的需求增加，同时也更愿意去游泳消暑。

因果推理所致力于解决的，是 “如果我们做了某事，会发生什么” 的问题。它通过合理的研究设计和科学的数据分析方法，像一位敏锐的侦探，努力识别出变量之间真正的因果关系。随着数据科学和机器学习的迅猛发展，因果推理在近年来受到了更为广泛的关注和应用。传统的机器学习方法主要侧重于预测和分类任务，例如预测股票价格的涨跌、对客户进行信用评级分类等。而因果推理则更加关注干预和决策，它能够深入分析在不同的干预措施下，结果会如何变化，从而为业务决策提供更具针对性和指导性的建议。

将因果推理技术融入信创方案，在证券行业的运营场景中具有不可忽视的重要意义。一方面，信创环境下的数据安全和自主可控为因果推理提供了坚实可靠的基础。在信创体系中，数据的存储、传输和处理都遵循严格的安全标准，确保数据的完整性和隐私性不受侵犯。这使得基于这些数据进行的因果分析结果更加可信，为决策提供了可靠的依据。其次，因果推理技术能够在复杂多变的市场环境中，如同精准的导航仪，识别和利用关键因素，制定出更加精准和有效的运营策略。通过这种深度融合，我们不仅能够提升业务决策的质量，使决策更加科学、

合理，还能够推动信创产业在金融科技领域的持续创新和发展，为金融科技的进步注入新的活力。

第二节 因果推理金融应用场景

在金融这个复杂的领域，因果推理有着丰富多样的应用场景。对于证券行业而言，准确地识别客户行为和市场变化之间的因果关系，对于制定有效的运营策略至关重要。

在客户的日常运营活动中，因果推理的应用体现在各个阶段，为券商的决策提供了有力的支持。

在客户获取阶段，券商面临着众多的营销渠道选择，如何确定哪些渠道最有效，是提高营销投入回报率的关键。因果推理在这里发挥了重要作用，它可以深入分析不同营销渠道对客户注册和开户的影响。例如，通过对大量数据的分析，研究传统的广告投放、社交媒体营销、合作伙伴关系以及口碑营销等各种渠道的效果。可能会发现，在某个特定地区，针对年轻投资者群体，社交媒体营销的效果显著，能够吸引大量潜在客户注册开户；而对于高净值客户，合作伙伴推荐和个性化的线下活动可能更具吸引力。通过这样的分析，券商可以确定最有效的营销渠道组合，避免资源的浪费，提高营销投入的回报率。

进一步地，在客户留存方面，客户对券商服务的满意度直接影响着他们的留存率。因果推理可以帮助券商深入研究客户对不同服务的反馈和其留存率之间的因果关系。例如，通过对客户服务响应时间的分析，可能发现如果客户咨询能够在 5 分钟内得到回复，客户的留存率会显著提高；而如果响应时间超过 15 分钟，客户流失的风险就会增加。此外，个性化投资建议的提供、客户界面的用户体验优化等方面也与客户留存率密切相关。通过这些深入的研究，券商可以针对性地优化服务内容和流程，提升客户体验，增强客户粘性。

在客户增长阶段，因果推理可以成为挖掘客户潜在价值的有力工具。它可以用来识别哪些产品特性或服务最能促进客户的交叉销售和升级。例如，通过对客户数据的详细分析，可能会发现某些客户群体对特定的投资组合产品有着较高的兴趣，或者对某些增值服务如专业的投资研究报告、定制化的投资策略等特别敏感。这些信息可以帮助券商设计定制化的营销活动，针对不同客户群体推荐合适

的产品和服务，从而提高客户的交叉购买率和升级率，最终提高客户生命周期价值。

在投资决策领域，因果推理同样发挥着重要作用。分析师可以借助因果推理分析宏观经济指标、行业动态等因素对股票价格的因果影响。例如，研究利率变动、通货膨胀率、行业政策调整等因素如何影响股票价格的走势。通过对这些因果关系的准确把握，分析师可以做出更准确的投资决策，为投资者提供更有价值的建议。

此外，在风险管理方面，因果推理可以帮助券商识别潜在的风险因素及其对投资组合的因果影响。例如，分析市场波动、信用风险、流动性风险等因素如何相互作用，影响投资组合的价值。通过提前识别这些因果关系，券商可以采取相应的措施进行风险防范，如调整投资组合的结构、设置风险预警指标等，确保投资组合的安全性和稳定性。

然而，尽管因果推理在金融领域具有巨大的应用潜力，但在证券行业中的广泛应用仍处于起步阶段。目前，大部分券商的运营决策主要基于经验和相关性分析，未能充分利用因果推理技术来挖掘深层次的因果关系。这种现状既反映了行业在技术应用上的滞后性，也为本研究提供了广阔的发展空间，激励着我们深入探索，为证券行业的发展带来新的突破。

第二章 因果推理的理论与原理

在深入探讨因果推理在信创方案及金融领域的应用之前, 有必要先对因果推理的理论与原理进行系统的梳理。因果推理致力于揭示变量之间的因果关联, 这一过程涉及诸多数学理论和方法, 它们为我们理解和分析因果关系提供了坚实的基础。

第一节 因果关系的基本概念

因果关系可简单理解为一个事件 (原因) 的发生必然导致另一个事件 (结果) 的发生。然而, 在实际的数据分析和研究中, 准确界定因果关系并非易事。从统计学角度来看, 变量之间的相关性是容易观察和度量的, 但相关性并不等同于因果关系。如前文提到的冰淇淋销量与游泳人数的例子, 它们之间存在相关性, 但并非因果关系。 在因果推理中, 我们引入了“干预”(Intervention) 的概念。干预是指对某个变量进行主动的操作或改变, 以观察其他变量随之产生的变化。如果通过干预变量 X, 能够稳定地引起变量 Y 的变化, 那么我们就可以认为 X 与 Y 之间存在因果关系。用数学语言表示, 可记为 $do(X = x)$, 表示对变量 X 进行干预, 使其取值为 x。

第二节 因果图模型

因果图模型 (Causal Graphical Models) 是一种用于表示变量之间因果关系的强大工具。它以图形的方式直观地展示了各个变量之间的因果联系, 为因果推理提供了清晰的框架。 因果图由节点和边组成, 节点代表变量, 边表示变量之间的因果关系。例如, 在一个简单的因果图中, 有变量 A、B 和 C, 如果存在从 A 到 B 的边, 以及从 B 到 C 的边, 那么表示 A 是 B 的原因, B 是 C 的原因, 即 $A \rightarrow B \rightarrow C$。 因果图模型中, 有几个重要的概念:

有向无环图 (Directed Acyclic Graph, DAG) : 因果图通常要求是有向无环图, 即图中不存在一条从某个节点出发, 经过一系列边后又回到该节点的路径。

祖先节点与后代节点: 在因果图中, 如果存在一条从节点 X 到节点 Y 的有

向路径，那么 X 是 Y 的祖先节点，Y 是 X 的后代节点。例如，在 $A \rightarrow B \rightarrow C$ 中，A 是 C 的祖先节点，C 是 A 的后代节点。

混杂因素（Confounder）：混杂因素是指同时影响原因变量和结果变量的其他变量，它可能会干扰我们对因果关系的判断。例如，在研究锻炼与健康之间的关系时，年龄可能就是一个混杂因素，因为年龄既可能影响人们锻炼的频率，也可能影响人们的健康状况。

第三节 因果效应的估计

因果效应（Causal Effect）是衡量原因变量对结果变量影响程度的指标。在因果推理中，准确估计因果效应是关键任务之一。 假设我们要研究变量 X 对变量 Y 的因果效应，常见的因果效应度量包括平均因果效应（Average Causal Effect, ACE) 和条件平均因果效应 (Conditional Average Causal Effect, CACE)。平均因果效应定义为：

$$ACE = E[Y|do(X = 1)] - E[Y|do(X = 0)]$$

其中，$E[Y|do(X = 1)]$ 表示对 X 进行干预，使其取值为 1 时，Y 的期望；$E[Y|do(X = 0)]$ 表示对 X 进行干预，使其取值为 0 时，Y 的期望。ACE 反映了在总体中，X 取值从 0 变为 1 时，Y 的平均变化量。 条件平均因果效应则是在给定某些协变量 Z 的条件下，X 对 Y 的因果效应，定义为：

$$CACE = E[Y|do(X = 1), Z = z] - E[Y|do(X = 0), Z = z]$$

其中，$Z = z$ 表示协变量 Z 取特定值 z。CACE 可以帮助我们更细致地分析在不同条件下，X 对 Y 的因果影响。为了估计因果效应，我们需要使用一些方法和技术。常用的方法包括随机对照试验（Randomized Controlled Trial，RCT）、倾向得分匹配（Propensity Score Matching，PSM）、工具变量法（Instrumental Variable Method）等。 随机对照试验是因果效应估计的黄金标准。在随机对照试验中，将研究对象随机分为实验组和对照组，对实验组进行干预，对照组不进行干预，然后比较两组的结果差异，从而估计因果效应。 倾向得分匹配是一种

在观察性研究中常用的方法。它通过计算每个个体接受干预的倾向得分（即个体接受干预的概率），然后将倾向得分相近的个体进行匹配，从而构建一个类似于随机对照试验的样本，进而估计因果效应。 工具变量法适用于存在内生性问题（即原因变量与误差项相关）的情况。通过引入一个与原因变量相关，但与误差项不相关的工具变量，我们可以利用工具变量来识别和估计因果效应。

第四节 因果推理的算法与技术

随着计算机技术和数据科学的发展，因果推理的算法与技术也得到了不断的创新和完善。一些基于机器学习的方法被广泛应用于因果推理领域，为我们处理复杂的数据和因果关系提供了有力的工具。贝叶斯网络（Bayesian Network）：贝叶斯网络是一种基于概率图模型的方法，它结合了概率论和图论的知识，能够有效地表示变量之间的不确定性和因果关系。贝叶斯网络通过节点表示变量，边表示变量之间的条件依赖关系，并使用条件概率表（Conditional Probability Table，CPT）来描述变量之间的概率关系。

结构因果模型（Structural Causal Model，SCM）：结构因果模型是一种更具结构性的因果推理框架，它将因果关系表示为一组结构方程。每个结构方程描述了一个变量如何依赖于其直接原因变量和外生变量。

因果发现算法（Causal Discovery Algorithms）：因果发现算法旨在从观测数据中自动发现变量之间的因果关系。这些算法通过分析数据中的统计规律和依赖关系，构建因果图模型。常见的因果发现算法包括 PC 算法、GES 算法等。PC 算法通过逐步删除不满足条件独立性的边来构建因果图；GES 算法则通过搜索最优的因果图结构，使得数据的似然性最大。 综上所述，因果推理的理论与原理涉及因果关系的基本概念、因果图模型、因果效应的估计以及因果推理的算法与技术等多个方面。这些知识为我们理解和应用因果推理提供了必要的基础，也为将因果推理融入信创方案以及在金融领域的广泛应用奠定了坚实的理论基石。

一、Tree - based 算法

基于树的算法以树结构为基础，通过对数据进行逐步划分来探寻变量之间的因果关系，以下是几种典型的基于树的因果推理算法：

提升树（Uplift Tree）：该方法运用基于树的算法，其分裂准则基于提升的差异。具体提出了三种量化分裂导致差异增益的方式，分别基于 KL 散度 (KL)、欧几里得距离（ED）和卡方（Chi）。这些方式通过衡量治疗组和对照组中感兴趣结果的概率分布差异，来决定树的分裂。例如，KL 散度通过特定公式计算不同叶子节点上治疗组和对照组样本均值的对数比例差异来衡量；欧几里得距离通过计算样本均值的平方差来衡量；卡方则通过计算样本均值差异的平方与对照组样本均值的比值来衡量。此外，还有基于ΔΔP 准则以及 IDDP 的提升随机森林算法，它们专门针对二叉树和二分类问题，分别依据特定的样本分裂准则来构建树结构。ΔΔP 准则通过计算不同分支响应率差异的差异来决定分裂；IDDP 则在ΔΔP 的基础上，结合特定的信息准则来确定分裂。另外，交互树（IT）通过最大化 统计量来确定样本分裂，以此挖掘变量间的交互作用；因果推断树 (CIT) 通过计算似然比检验统计量来进行样本分裂；基于情境化治疗选择的提升随机森林（CTS）则依据情境因素，通过特定的样本分裂准则来实现治疗选择和因果效应估计。

二、Meta - learner 算法

Meta - learner 算法是一类基于已有多个基础学习器的结果进行二次学习的算法，旨在更高效、准确地估计因果效应。常见的 Meta - learner 算法如下：

S - Learner：该算法使用单个机器学习模型来估计治疗效果。首先，结合协变量 X 和治疗指示变量 W，通过公式 $\mu(x,w) = E[Y \mid X = x, W = w]$ 估计平均结果 μ_x。然后，将条件平均治疗效果（CATE）估计定义为

$\hat{\tau}(x) = \hat{\mu}(x, W = 1) - \hat{\mu}(x, W = 0)$。在模型中纳入倾向得分有助于减少由正则化引起的混杂偏差。当对照组和治疗组在协变量方面存在较大差异时，单个线性模型可能难以充分处理这种情况。

T - Learner：由两个阶段构成。第一阶段，利用机器学习模型分别估计对照组和治疗组的平均结果$\mu_0(x)$和$\mu_1(x)$，公式为$\mu_0(x) = E[Y(0)|X = x]$ $\mu_1(x) = E[Y(1)|X = x]$。第二阶段，将 CATE 估计定义为$\hat{\tau}(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x)$。

X - Learner：作为 T - learner 的扩展，包含三个阶段。第一阶段与 T - learner 相同，即估计$\mu_0(x)$和$\mu_1(x)$。第二阶段，基于$\mu_0(x)$和$\mu_1(x)$推算治疗组和对照组的用户级治疗效果，并进一步估计$\tau_1(x)$和$\tau_0(x)$。第三阶段，通过$\tau_1(x)$和$\tau_0(x)$的加权平均来定义 CATE 估计，公式为$\tau(x) = g(x)\tau_0(x) + (1 - g(x))\tau_1(x)$，其中 g ∈ (0,1]，通常可使用倾向得分来确定 。

R - Learner：运用交叉验证的折叠外估计结果$\hat{m}^{(-i)}(x_i)$和倾向得分$\hat{e}^{(-i)}(x_i)$。第一阶段，通过交叉验证的机器学习模型拟合$\hat{m}(x)$ 和$\hat{e}(x)$。第二阶段，通过最小化 R - 损失$\hat{L}_n(\tau(x))$来估计治疗效果，公式为:

$$\hat{L}_n(\tau(x)) = \frac{1}{n}\sum_{i=1}^{n}\left((Y_i - \hat{m}^{(-i)}(X_i)) - (W_i - \hat{e}^{(-i)}(X_i))\tau(X_i)\right)^2$$

双稳健（Doubly Robust，DR）学习器：通过交叉拟合双稳健得分函数来估计 CATE。首先将数据随机分为 3 个分区，第一阶段，分别拟合倾向得分模型和结果回归模型。第二阶段，从伪结果中拟合 CATE 模型，伪结果公式为

$$\phi = \frac{W - \hat{e}(X)}{\hat{e}(X)(1 - \hat{e}(X))}\left(Y - \hat{m}_W(X)\right) + \hat{m}_1(X) - \hat{m}_0(X)$$

第三阶段，重复前两个阶段两次，最终取三个 CATE 模型的平均值作为最终结果。

三、工具变量算法

两阶段最小二乘法（2 - Stage Least Squares，2SLS）：在因果推理中，当存在内生性问题，即原因变量 与误差项相关时，为了准确估计 对结果变量的因果效应，引入工具变量 。2SLS 算法分为两个阶段，第一阶段，将原因变量 对工具变量进行回归，得到预测值；第二阶段，把 对第一阶段得到的 的预测值进行回归，从而得到因果效应的估计值。这种方法通过利用工具变量与原因变量的相关性以及与误差项的不相关性，有效地解决了内生性问题，在许多领域都有广泛应用。

双稳健工具变量（Doubly Robust Instrumental Variable，DRIV）学习器：结合了双稳健方法和工具变量的思想。在处理内生性问题的同时，通过随机将数据分为 3 个分区，在不同分区上分别拟合倾向得分模型、结果回归模型，并利用这些模型从特定的损失函数中拟合条件局部平均治疗效果模型。经过多次重复拟合和计算，最终取三个条件局部平均治疗效果模型的平均值作为结果，提高了估计的稳健性和准确性。

四、基于神经网络的算法

CEVAE（Counterfactual Embeddings for Variational Autoencoders）：该算法融合了变分自编码器（VAE）和反事实推理的思想。它通过构建变分自编码器，将数据映射到低维的潜在空间，在这个空间中进行反事实推理，从而估计因果效应。这种方法能够有效处理高维数据和复杂的因果关系，为因果推理提供了新的思路和方法。

DragonNet：是一种基于神经网络的因果推理模型，通过构建多个子网络来分别估计不同处理组的结果分布。该模型能够自动学习数据中的复杂模式和因果关系，在处理高维数据和多处理组问题时表现出良好的性能，能够较为准确地估计因果效应。

五、治疗优化算法

反事实单元选择（Counterfactual Unit Selection）：该方法基于反事实逻辑来选择接受治疗的单元。通过考虑不同类型的个体（如依从者、总是接受者、从不接受者、违抗者）在不同治疗情况下的收益，利用反事实逻辑构建优化问题，从而选择最有可能从治疗中受益的个体。具体通过估计不同个体特征下属于各类人群的概率，来确定最优的治疗选择。

反事实价值估计器（Counterfactual Value Estimator）：利用标准机器学习模型预测一个单元在不同治疗条件下的结果，然后根据结果计算在特定治疗下的预期价值。该预期价值考虑了治疗的成本以及有利事件发生的概率等因素，通过公式 $\mathbb{E}[(v - cc_w)Y_w - ic_w]$ 进行计算，其中 Y_w 是在给定治疗 w 下有利事件发生的概率，v 是有利事件的价值，cc_w 是有利事件发生时治疗的成本，ic_w 是无论结果如何治疗所产生的成本。这种方法为决策提供了关于不同治疗方案预期价值的参考，有助于选择最优的治疗策略。

第三章 因果推理算法自研

第一节 因果推理建模

本研究基于现有因果推理理论和方法，结合证券行业的特点和需求，自研了一套因果推理算法。该算法的流程主要包括以下几个步骤：

一、数据收集与预处理

首先，从数据系统中收集大量的数据，包含了客户基本信息、交易记录、行为数据等多方面的信息。然而，这些数据具有多源、异构、海量的特点，需要进行预处理，才能展现出其价值。

预处理过程包括数据清洗、特征工程、数据集成等一系列复杂而重要的操作。数据清洗是为了去除数据中的噪声和错误，填充缺失值，确保数据的准确性和完整性。例如，在客户基本信息中，可能存在一些缺失的字段，如年龄、职业等，

需要通过合理的方法进行填充。可以利用统计学方法，根据其他客户的信息进行估算，或者采用机器学习算法进行预测填充。

特征工程则是从原始数据中提取有价值的特征，将数据转化为适合模型处理的形式。在本项目的模型中，我们采用了基础信息类、资产与交易类、浏览天数类、点击次数类以及末次点击类等特征。对于基础信息类特征，可能需要对一些分类变量进行编码，将其转化为数值型变量，以便模型能够更好地处理。对于资产与交易类特征，除了进行数据清洗外，还需要进行分箱处理，将连续的数值型变量划分为不同的区间，以减少数据的波动性，提高模型的稳定性。

数据集成是将来自不同数据源的数据整合到一起，确保数据的一致性和连贯性。例如，将客户在不同交易平台上的交易记录进行整合，统一数据格式和标准，为后续的分析提供可靠的数据基础。

二、因果模型构建

根据研究问题和数据特点，选择合适的因果模型是构建因果推理算法的关键环节。在本研究中，主要采用了基于倾向性评分匹配（PSM）和 meta - learner 的方法。

倾向性评分匹配是一种常用的因果推理方法，它通过计算每个样本的倾向性评分，将处理组和对照组中的样本进行匹配，从而消除混杂因素的影响，估计因果效应。具体来说，倾向性评分是指在给定协变量的情况下，个体接受处理的概率。通过计算倾向性评分，可以找到与处理组样本在协变量上相似的对照组样本，使得处理组和对照组在混杂因素上具有可比性。这样，就可以通过比较处理组和对照组的结果，来估计处理的因果效应。

Meta - learner 则是一种基于机器学习的因果推理方法，它能够处理高维数据和复杂的非线性关系，提高因果效应估计的准确性。Meta - learner 利用机器学习算法来学习因果效应的估计函数，通过对大量数据的学习，能够自动捕捉数据中的复杂模式和关系。与传统的因果推理方法相比，Meta - learner 具有更强的适应性和灵活性，能够更好地应对实际应用中的各种挑战。

在构建因果模型的过程中，还需要考虑运营人群四象限的问题。模型的目标

是识别出某类活动的运营敏感人群，对于其他的三种人群，在当前活动中无需给予特别的关注。通过对不同人群的精准识别，可以更有针对性地制定运营策略，提高运营效率。

三、运营人群四象限

图 1 运营人群四象限

模型的目标是识别出某类活动的运营敏感人群，对于其他的三种人群，在当前活动中无需给予特别的关注。

四、数据分析

在进行数据分析时，首先是对特征的深入分析。在本项目的模型中，我们精心选取了 5 个类别共 51 个特征。这些特征涵盖了客户的多个方面信息，为我们深入了解客户行为和市场变化提供了丰富的数据支持。

出于对训练资源等的综合考虑，我们选择 2024 年的 4 月、5 月、6 月、7 月、8 月、9 月的特征，依次对应下个月的转化，作为训练数据。10 月的特征，11 月的转化，作为验证数据。

在三个不同业务场景中，均采用分层抽样策略构建训练数据集，从大规模负样本中按特定比例筛选数据并与全量正样本共同构成模型输入，另有独立于训练集的验证集用于模型效果评估。各场景数据分布存在差异，部分场景正样本占比相对均衡，部分场景正样本稀缺性显著需采用特殊采样策略，部分场景正样本密度较高需关注分类精度优化。

所有场景均通过负样本降采样控制训练资源消耗，在保证数据代表性的前提

下实现训练效率与模型性能的平衡。

五、模型训练与验证

使用预处理后的数据对因果模型进行训练，这是一个不断优化和调整的过程。在训练过程中，通过调整模型参数，如学习率、正则化参数等，来优化模型性能，使模型能够更好地拟合数据，提高预测的准确性。

为了确保模型的泛化能力和稳定性，我们采用交叉验证等方法对模型进行评估。交叉验证是将数据集划分为多个子集，每次使用其中一个子集作为验证集，其余子集作为训练集，进行多次训练和验证。通过这种方式，可以更全面地评估模型的性能，避免模型过拟合或欠拟合的问题。

同时，通过比较不同模型的性能指标，如准确率、召回率、F1 值等，选择最优的模型进行后续的分析。目前建模的场景包括三种增值工具的运营活动投放。通过对不同场景下的模型进行训练和验证，我们能够针对不同的业务需求，选择最合适的模型，为运营决策提供更准确的支持。

六、因果效应估计

利用训练好的因果模型，对感兴趣的变量之间的因果关系进行估计，这是因果推理的核心任务。例如，分析某种运营策略对客户活跃度的因果效应，或者分析宏观经济因素对股票价格的因果影响等。

在估计因果效应时，需要考虑到各种可能的混杂因素和偏差，采用合适的方法进行校正和调整，以获得准确的因果效应估计值。例如，在分析运营策略对客户活跃度的影响时，可能存在一些混杂因素，如客户的年龄、投资经验、市场环境等，这些因素可能会同时影响运营策略的实施效果和客户的活跃度。因此，需要通过合理的方法，如倾向性评分匹配、工具变量法等，来控制这些混杂因素的影响，确保因果效应估计的准确性。

对因果效应估计结果进行深入分析和解释，理解变量之间的因果关系及其背后的机制。这就像是揭开一层神秘的面纱，让我们能够清晰地看到事物之间的内在联系。通过可视化等手段，将因果关系以直观的方式呈现出来，为业务决策提供清晰的依据。例如，可以使用因果图、柱状图、折线图等方式，将因果效应的

大小、方向等信息直观地展示出来，使决策者能够一目了然。

同时，结合实际业务场景，对因果分析结果进行合理性验证，确保结果的可靠性和实用性。例如，在分析某种运营策略对客户投资回报率的影响时，需要将分析结果与实际的市场情况、客户反馈等进行对比，验证结果是否符合实际情况。如果结果与实际情况不符，需要进一步分析原因，对模型进行调整和优化，直到得到可靠的结果。

第四章 成果展示

第一节 特征处理

我们精心打造了一个高效的筛选体系，对客户画像指标进行精细化筛选，并据此成功开发了一套客户群体用户画像分析系统。该系统依托于大数据和人工智能技术，深入挖掘海量客户数据，精准捕捉客户的特征、需求和行为模式。

例如，系统通过综合分析客户的特征，生成详尽的用户画像报告，为制定针对性的运营策略和服务提供了坚实的数据基础。

在券商的用户画像构建过程中，我们注意到了以下几个特点：首先，数据稀疏性问题显著，大多数特征的缺失率超过 90%，即存在大量空白数据。为了降低这一问题对特征提取的影响，我们在数据处理阶段仅考虑了高活跃用户。其次，数据跨度区间较大，特别是在某些数据中，存在许多离群值。对于这些变量，我们实施了二次分箱处理，以减少离群值对模型拟合的不利影响。

针对不同特征，我们采取了差异化的数据处理策略。对于基础信息类特征，我们进行了数据清洗，包括填充缺失值和去除 NaN 值；对于资产与交易类特征，除了数据清洗外，还进行了分箱处理；对于浏览天数类特征，我们同样进行了数据清洗；对于点击次数类特征，除了数据清洗外，还进行了分箱处理；而对于末次点击类特征，我们在数据清洗的基础上，将其处理为距离特定日期的天数（数字），并进行了分箱处理。

第二节 因果推理

提出并实现了基于因果推理的运营方案，为业务决策提供了科学依据。通过深入分析数据并识别因果关系，该方案能够帮助券商优化运营策略，提高运营效果。

在实际应用中，利用因果推理算法分析了不同运营活动对客户留存率、活跃度和投资回报率的因果影响。例如，通过倾向性评分匹配和因果模型等方法，发现针对特定客户群体的个性化推荐活动能够显著提高客户的投资回报率，而定期举办的线上投资讲座则对提高客户的活跃度有积极作用。基于这些分析结果，调

整运营方案。

在验证集上的结果如下图所示

在场景 A 中，验证集模型预测各层的命中率如下表所示。

表 1 场景 A 预测值分层对应命中率

验证集模型预测结果分层	命中率
90.00% ~ 100.00%	69.95%
80.00% ~ 90.00%	16.26%
70.00% ~ 80.00%	5.42%
60.00% ~ 70.00%	3.45%
50.00% ~ 60.00%	1.97%
40.00% ~ 50.00%	0.99%
30.00% ~ 40.00%	0.49%
20.00% ~ 30.00%	0.49%
10.00% ~ 20.00%	0.49%
0.00% ~ 10.00%	0.49%

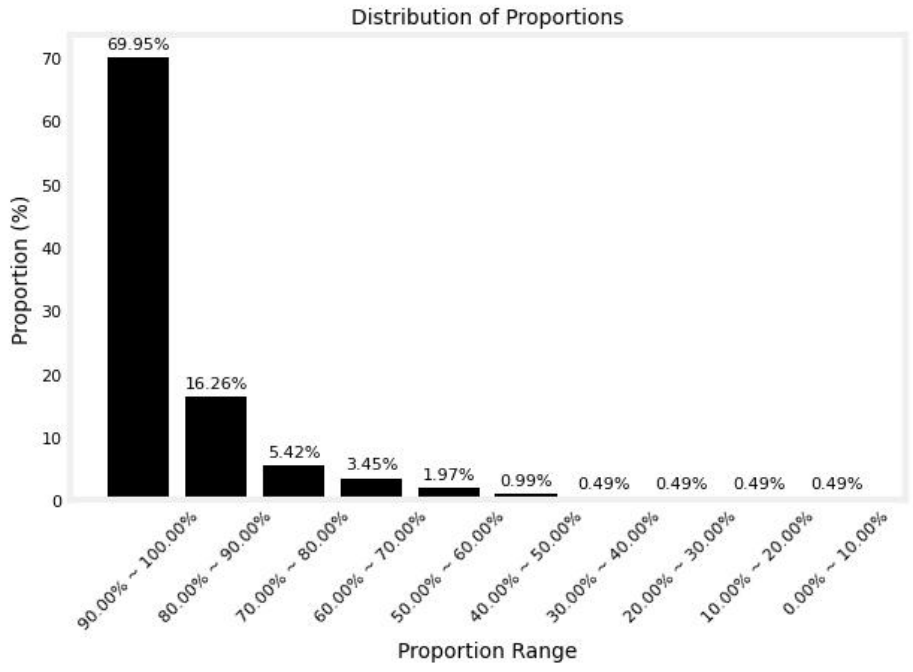

图 2 场景 A 预测值分层对应命中率柱状图

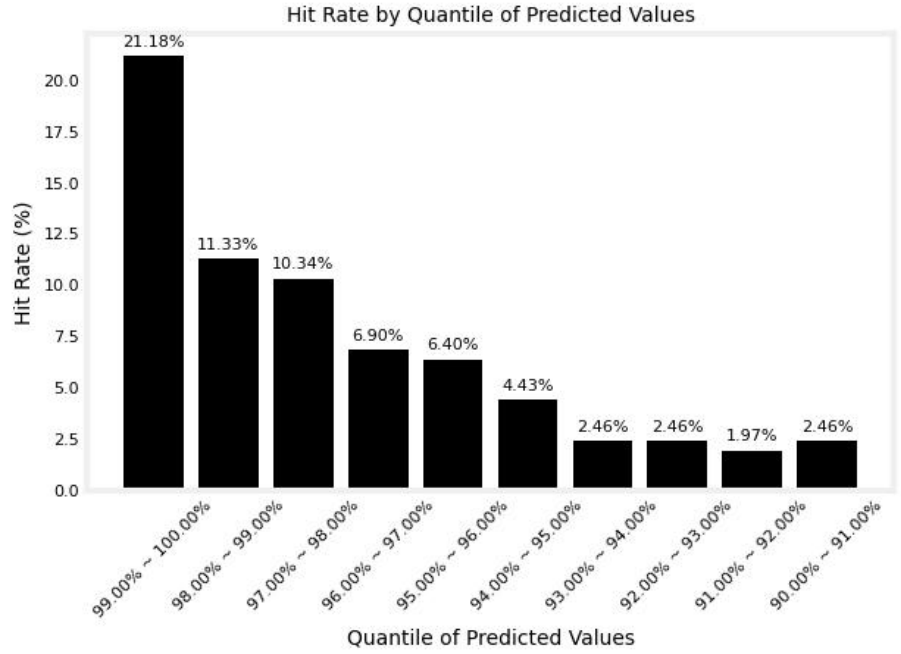

图 3 场景 A 预测值分层对应命中率柱状图（前 10%）

在场景 B 中，验证集模型预测各层的命中率如下表所示。

表 2 场景 B 预测值分层对应命中率

验证集模型预测结果分层	命中率
90.00% ~ 100.00%	74.00%
80.00% ~ 90.00%	12.50%
70.00% ~ 80.00%	4.25%
60.00% ~ 70.00%	4.00%
50.00% ~ 60.00%	0.75%
40.00% ~ 50.00%	1.50%
30.00% ~ 40.00%	1.00%
20.00% ~ 30.00%	0.75%
10.00% ~ 20.00%	0.00%
0.00% ~ 10.00%	1.25%

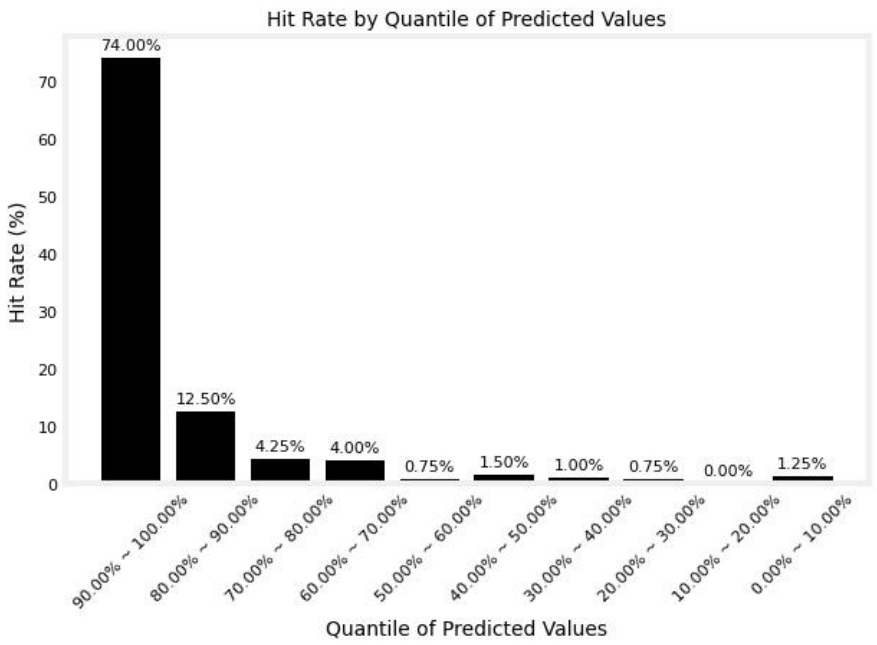

图 4 场景 B 预测值分层对应命中率柱状图

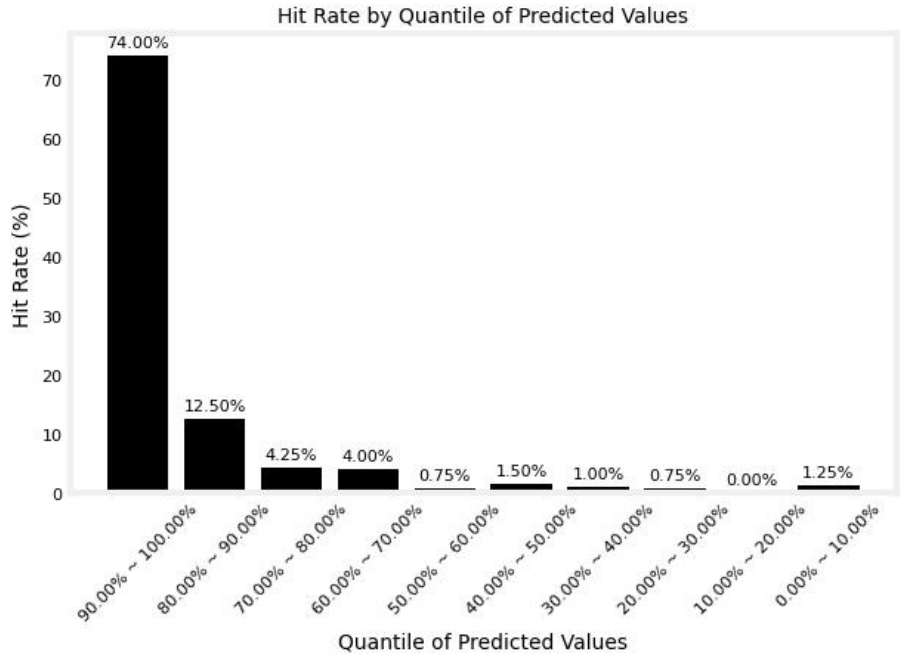

图 5 场景 B 预测值分层对应命中率柱状图（前 10%）

在场景 C 中，验证集模型预测各层的命中率如下表所示。

表 3 场景 C 预测值分层对应命中率

验证集模型预测结果分层	命中率
90.00% ~ 100.00%	69.88%
80.00% ~ 90.00%	16.87%
70.00% ~ 80.00%	3.61%
60.00% ~ 70.00%	3.61%
50.00% ~ 60.00%	2.41%
40.00% ~ 50.00%	0.00%
30.00% ~ 40.00%	1.20%
20.00% ~ 30.00%	1.20%
10.00% ~ 20.00%	0.00%
0.00% ~ 10.00%	1.20%

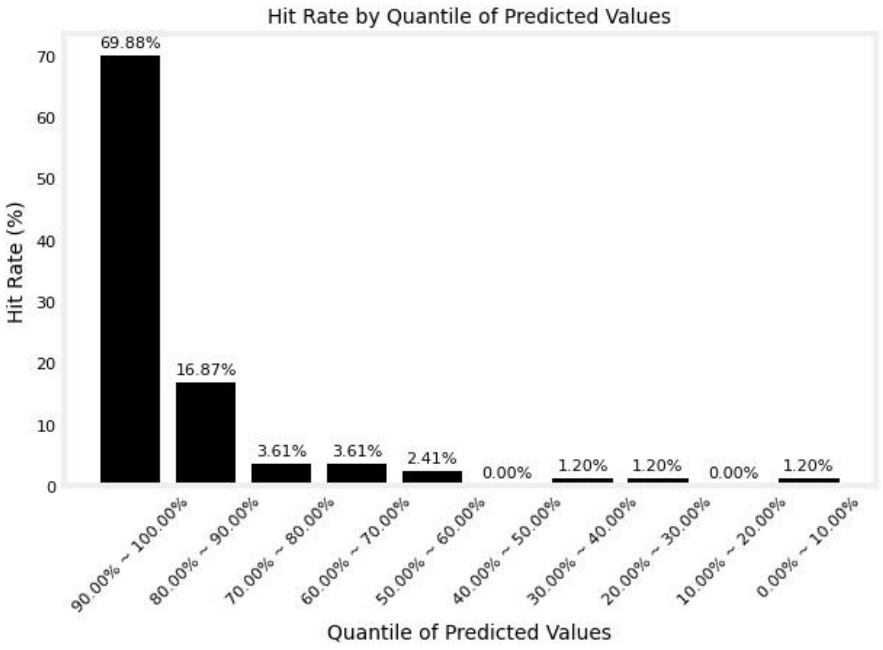

图 6 场景 C 预测值分层对应命中率柱状图

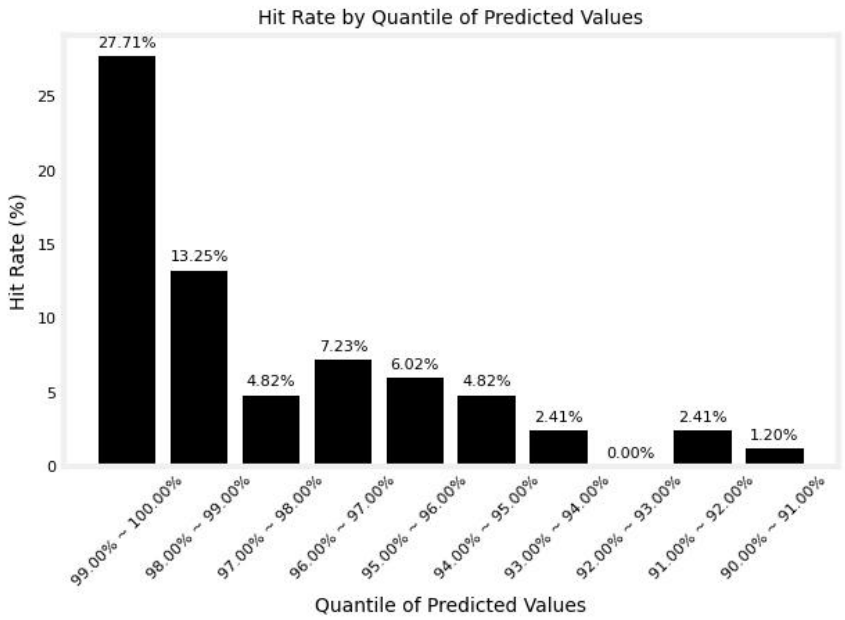

图 7 场景 C 预测值分层对应命中率柱状图（前 10%）

在三个场景中验证集的排名前 20%样本命中真实正样本共占比 85%+，达到了业务对线上模型的要求。

第三节 全栈信创智能运营策略引擎

设计和实现了一套完整的软硬件全栈信创智能运营策略引擎。该引擎涵盖了数据采集、数据处理、智能算法、决策支持系统等多个方面的功能，能够精准建立用户画像、智能圈选目标客户，并基于因果推理、机器学习、深度学习等先进技术手段实现智能决策支持。

在核心功能模块方面，对基于因果推理的智能决策支持算法进行了充分优化和调试，确保其在实际场景中具备较高的性能和稳定性。例如，在处理大规模客户数据时，算法能够快速准确地估计因果效应，为运营决策提供及时的支持。同时，该引擎还具备良好的可扩展性和兼容性，能够适应证券行业不断变化的业务需求。

第五章 课题研究总结

本课题围绕证券行业软硬件全栈信创智能运营策略引擎展开研究，取得了一系列有价值的成果。通过引入因果推理技术，填补了证券行业在这一领域落地应用的空白，为证券行业的运营决策提供了更加科学、精准的方法和工具。

在研究过程中，我们深入研究了因果推理的理论和方法，并结合证券行业的实际需求，自研了一套适合证券行业的因果推理算法。通过该算法，我们成功地解决了用户画像分析、运营方案制定、智能运营策略引擎设计等关键问题，实现了对客户的精准圈选和识别，提高了客户运营活动的效率。

同时，我们还设计了一系列具有针对性的运营活动策划方案和运营投放策略，通过实际应用验证了这些方案和策略的有效性和可行性。这些成果不仅为券商带来了实际的业务价值，也为证券行业的数字化转型和创新发展提供了有益的借鉴。

然而，本研究也存在一些不足之处。例如，在因果推理算法的准确性和效率方面，仍有进一步提升的空间；在与实际业务的深度融合方面，还需要不断探索和优化。未来的研究可以从以下几个方面展开：一是进一步优化因果推理算法，提高其在大规模数据和复杂业务场景下的性能和准确性；二是加强与其他人工智能技术的融合，如深度学习、强化学习等，实现更智能化的运营决策；三是深入研究证券行业的新业务和新需求，不断拓展智能运营策略引擎的应用场景和功能。

总之，本课题的研究成果为证券行业的发展提供了新的思路和方法，具有重要的理论意义和实际应用价值。我们相信，随着技术的不断进步和研究的深入开展，因果推理技术在证券行业将有更广泛的应用前景，为证券行业的创新发展注入新的动力。