卿诗佳-23212129
邓杰-23212116

基于 LightGBM 模型的银行产品客户认购预测

引言

在数字经济时代，金融服务业的竞争不断加剧，银行业也不例外。为了适应市场的变化，银行需要不断调整其业务策略和营销模式。然而，随着银行产品种类的日益丰富和复杂性增加，传统的营销模式面临诸多挑战。例如，银行需要通过营销活动推广多种产品，但不同客户对同一产品的需求差异可能很大。如何在海量客户中高效识别潜在购买者成为银行产品推广中的关键难题。现有的营销方式通常基于经验和通用规则，但这种方式无法有效利用客户数据和行为模式，既浪费了资源，也难以实现个性化的营销。

为了解决这一问题，银行开始引入数据驱动的精准营销方式，即通过分析客户行为数据，预测客户是否会对某些产品产生购买兴趣。精准营销的概念最早由 Peppers 和 Rogers（1997）提出，其核心在于基于客户数据构建个性化营销策略。近年来，精准营销逐渐被应用于金融行业，特别是银行领域。例如，许多银行开始使用客户的交易记录、贷款历史和社交网络行为来预测客户的潜在需求（Günther et al., 2014）。相比传统的基于规则的营销方式，精准营销能够更好地满足客户需求，同时降低银行的运营成本。预测结果可以帮助银行更高效地分配营销资源，优化推广策略。例如，预测模型可以识别出可能对某一款理财产品感兴趣的客户，从而促使银行将相关的产品信息推荐给这些客户，减少不必要的营销成本。此外，通过分析哪些客户可能流失，银行可以及时采取挽留措施，从而提高客户黏性和忠诚度。近年来，随着大数据和机器学习技术的快速发展，金融机构已经能够利用丰富的客户历史数据和行为特征，构建高效的预测模型，从而提升营销活动的整体效能。

然而，在实际应用中，精准营销面临一些挑战，如数据的高维性和复杂性，由于银行客户数据通常包含数百个特征，如何有效利用这些特征进行预测是一个难题。其次，客户行为的非线性和随机性也是一大难题，客户的购买行为受多种因素影响，包括宏观经济环境、个体偏好和社会网络效应，这些因素之间往往存在复杂的交互关系。但是近年来，随着大数据技术和机器学习算法的快速发展，通过构建基于历史数据的分类模型，机器学习算法可以识别隐藏在数据中的模式和规律，从而对未来的客户行为进行预测。机器学习技术因其强大的数据分析能力，被广泛应用于金融行业的多个领域（Bello, 2023）。常见的应用包括信用风险评估、欺诈检测、股票市场预测和客户行为分析等。在银行客户购买预测中，机器学习的优势在于其能够处理复杂的非线性关系，并自动从数据中学习模式。

所以，本研究旨在探索如何通过机器学习技术预测客户的银行产品购买行为。我们利用来自阿里云天池平台的 30,000 条银行客户数据，构建了多种分类模型，包括决策树、随机森林、LightGBM 和 XGBoost 等主流算法。这些算法被广泛应用于分类任务中，并以其各自的特点和优势适应不同的场景。为了进一步提升模型的预测性能，我们还采用了堆叠集成方法，将表现较优的随机森林和 LightGBM 模型进行组合，并通过逻辑回归作为元学习器进行整合。

在模型性能评估方面，我们采用了 AUC（Area Under the ROC Curve）作为核心指标，并结合 F1 分数辅助评价不同模型的预测能力。研究结果表明，单一模型中，随机森林和 LightGBM 表现最佳，AUC 分别达到 0.8896 和 0.8871，而堆叠模型的 AUC 提升至 0.8917。尽管堆叠模型具有更高的预测能力，但其复杂性和部署成本较高。基于实际应用的考虑，我们最终选择了 LightGBM 模型作为部署模型。

此外，我们对 LightGBM 模型的特征重要性进行了可视化分析，揭示了影响客户购买行为的关键因素，包括客户与银行交互的持续时间（duration）、宏观经济指标（emp_var_rate）以及历史联系间隔（pdays）等。这些发现不仅有助于理解客户行为模式，还为银行在产品推广策略上提供了具体的优化建议。

数据与方法

1. 数据预处理

1.1 数据来源

本研究使用的 30,000 条银行客户数据来自阿里云天池平台，数据包含了 18 个特征和目标变量 subscribe，其中目标变量表示客户是否认购银行产品（“yes”表示购买，“no”表示未购买）。数据特征既包括类别型特征（如客户婚姻状态 marital、住房贷款情况 housing 等），也包括数值型特征（如客户年龄 age、营销活动期间的联系次数 campaign 等）。完整的数据特征描述见表 1。

字段	说明	字段	说明
age	年龄	job	职业
marital	婚姻	default	信用卡是否有违约
housing	是否有房贷: yes or no	contact	联系方式
month	上一次联系的月份	day_of_week	上一次联系的星期几
duration	上一次联系的时长（秒）	campaign	活动期间联系客户的次数
pdays	上一次与客户联系后的间隔天数	previous	本次营销活动前，与客户联系的次数
poutcome	之前营销活动的结果	emp_var_rate	就业变动率（季度指标）
cons_price_index	消费者价格指数（月度指标）	cons_conf_index	消费者信心指数（月度指标）
lending_rate3m	银行同业拆借率（每日指标）	nr _employed	雇员人数（季度指标）
subscribe*	客户是否进行购买：yes或no		

表 1. 数据特征表

1.2 数据清理

训练集和测试集中均没有缺失值。所有特征在两个数据集中的记录数都是完整的，不需要填补缺失值。描述性统计发现各数值特征均没有明显异常。

1.3 特征编码与标准化

目标变量编码：我们使用 LabelEncoder 将目标变量 Subscribe 编码为二分类数值型数据：yes 编码为 1，no 编码为 0。

类别特征编码：使用 OrdinalEncoder 对所有类别型特征进行数值编码，将字符串转化为可输入模型的数值。

数值型特征标准化：使用 StandardScaler 分别对数值型特征进行标准化，处理后的数据均值为 0，标准差为 1，以提升模型的收敛速度及性能。

1.4 数据集划分

将数据划分为训练集和测试集。在训练集上使用 5 折交叉验证评估模型性能。

2. 模型构建与评估

2.1 决策树模型

决策树（Decision Tree）是一种广泛应用于监督学习的分类模型，其主要思想是基于特征的分裂，将数据逐步划分为不同的子集，从而形成一棵树状结构，最终通过树叶节点的类别标签进行预测。决策树模型以其直观的结构和较高的可解释性著称，尤其适用于结构化数据的分析。在分类任务中，决策树会评估各个特征的分裂点，通过选择能够最大化信息增益或最小化基尼指数的分裂点，依次进行特征分裂。这一过程的最终目标是使每个分裂后生成的子集尽可能纯净，即使得同一子集中样本的类别一致性最高。

为了提升决策树模型的性能，我们利用了 Optuna 框架对模型的超参数进行了自动化调优。调优的目标是通过交叉验证最大化 F1 分数，以在精确率（Precision）和召回率（Recall）之间取得平衡。F1 分数是分类任务中常用的综合指标，特别适用于类别不平衡的场景。在 200 次超参数搜索中，最终选定的最优参数组合为：max_depth=15, min_samples_split=8, min_samples_leaf=9, criterion=gini。

在获得最优超参数后，我们使用训练集和测试集分别评估了决策树模型的性能。在训练集上的准确率为 92.45%，而测试集的准确率为 86.49%，表明模型能够较好地拟合训练数据，但在未见数据上的性能有所下降。此外，训练集的 F1 分数为 0.6726，而测试集的 F1 分数

仅为 0.4381。这反映出模型在测试数据上的精确率和召回率之间未能达到良好的平衡。

2.2 随机森林模型

随机森林（Random Forest）是一种基于决策树的集成学习方法，由 Breiman 于 2001 年提出，通过构建多棵独立的决策树并将其预测结果进行集成，从而提高整体模型的分类或回归性能。在分类任务中，随机森林会对多个决策树的分类结果进行投票（多数投票法），以确定最终分类；在回归任务中则通过平均每棵树的预测值得到最终输出。随机森林通过在训练数据和特征选择上引入随机性，显著降低了单一决策树模型过拟合的风险，因而在泛化能力和稳定性方面具有突出表现。

同样地，为了优化随机森林的性能，我们使用了 Optuna 框架对超参数进行了自动化调优。在 200 次超参数搜索中，最终选定的最优参数组合为：n_estimators=125, max_depth=20, min_samples_split=4, min_samples_leaf=1, criterion=gini。

在获得最优超参数后，我们使用训练集和测试集对随机森林模型进行了性能评估。结果显示，随机森林在训练集上的准确率为 99.38%，而测试集的准确率为 88.36%。在 F1 分数方面，训练集的 F1 分数为 0.9757，表明模型在训练数据上的精确率和召回率之间达到了几乎完美的平衡。然而，测试集的 F1 分数为 0.4229，与训练集相比有显著下降，表明模型在泛化能力上有一定不足。

2.3 XGBoost 模型

XGBoost（Extreme Gradient Boosting）是一种基于梯度提升框架的增强型模型，由 Chen 和 Guestrin 于 2016 年提出。作为一种高效且灵活的机器学习算法，XGBoost 具有强大的非线性建模能力，广泛应用于分类和回归任务中。XGBoost 通过逐步优化目标函数（如对数损失 logloss）提高模型性能，其独特的优势包括支持并行计算、提供正则化项以防止过拟合，以及对缺失值具有内在的鲁棒性。XGBoost 的核心原理是基于加法模型逐步构建多棵决策树，每棵树都试图最小化前一棵树的预测误差，从而提升整体模型的准确性和稳定性。这种逐步优化的过程使其能够很好地捕捉数据中的复杂模式，并在各种数据集上表现优异。

为了优化 XGBoost 的性能，我们使用了 Optuna 框架对超参数进行了自动化调优。调优的目标是最小化 logloss，即对数损失函数，logloss 是衡量预测概率与真实值之间差距的重要指标，越小表示模型的预测越接近真实值。通过定义超参数的搜索空间，包括树的深度 m

ax_depth、学习率 learning_rate、树的数量 n_estimators、以及数据和特征采样比例 subsample 和 colsample_bytree，我们在 200 次超参数搜索中找到了最佳参数组合：max_depth=7，learning_rate=0.0257，n_estimators=200，subsample=0.7736，colsample_bytree=0.9101。此外，我们使用了早停法（early_stopping_rounds=10）在验证集上动态调整训练轮次，确保模型在 logloss 不再下降时及时停止训练，以避免过拟合。

在获得最优超参数后，我们使用训练集和测试集对 XGBoost 模型进行了性能评估。结果显示，XGBoost 在训练集上的准确率为 93.52%，在测试集上的准确率为 88.69%。训练集的高准确率表明模型能够充分学习数据中的复杂模式，而测试集的表现则表明模型具备一定的泛化能力。在 F1 分数方面，训练集的 F1 分数为 0.6910，表明模型在训练数据上精确率和召回率之间取得了良好的平衡。然而，测试集的 F1 分数为 0.4521，较训练集有所下降，反映出模型在未见数据上的表现仍有改进空间。

2.4 LightGBM 模型

LightGBM（Light Gradient Boosting Machine）是一种基于梯度提升决策树（GBDT）的高效分类模型。作为一种专为处理大规模数据和高维特征而设计的机器学习算法，LightGBM 采用基于叶节点的分裂策略（Leaf-wise Growth）和直方加速算法（Histogram-based Algorithm），显著提升了训练效率，同时保持了较高的预测性能。与传统的 GBDT 方法相比，LightGBM 在训练速度和内存占用方面具备显著优势，并且能够更好地适应稀疏数据和类别不平衡场景。此外，LightGBM 支持分布式训练和特征重要性分析，能够为大规模数据的分类、回归和排序任务提供高效且易解释的解决方案。

为了优化 LightGBM 模型的性能，我们使用 Optuna 框架对超参数进行了自动化调优，目标是最大化 F1 分数，以在精确率（Precision）和召回率（Recall）之间取得平衡。F1 分数是分类任务中综合性能的关键指标，特别适用于目标变量类别分布不平衡的场景。调优的关键参数包括：num_leaves（树的最大叶节点数，控制模型复杂度）、learning_rate（学习率，影响模型的收敛速度）、n_estimators（基学习器数量）、subsample（数据子样本比例，用于减少过拟合）和 colsample_bytree（特征子样本比例，用于增强泛化能力）。经过 200 次超参数搜索，最终选定的最优参数组合为：num_leaves=176，learning_rate=0.0299，n_estimators=138，subsample=0.7823，colsample_bytree=0.9831。

在获得最优超参数后，我们对 LightGBM 模型进行了训练和测试。结果表明，LightGB

M 在训练集上的准确率为 97.69%，而测试集的准确率为 88.27%，说明模型对训练数据的拟合能力较强，同时在未见数据上的泛化性能也较为出色。在 F1 分数方面，训练集的 F1 分数为 0.6726，而测试集的 F1 分数为 0.4465，表明模型在训练数据和测试数据上的性能较为稳定。

2.5 交叉验证

交叉验证（Cross-Validation）是一种用于评估模型性能和稳定性的重要技术，其核心思想是通过多次将数据划分为训练集和验证集，减少因数据划分方式导致的评估结果偏差，从而更全面地衡量模型的泛化能力。在实际任务中，单次的训练测试划分可能导致模型评估结果的偶然性，因为某些特定的数据划分可能无法代表总体分布。而通过交叉验证，可以更充分地利用数据，同时减小评估结果的波动，从而为模型的可靠性和适用性提供更高的保障。

本研究采用了 5 折交叉验证（5-Fold Cross-Validation）来评估模型性能。具体而言，我们将训练数据随机分为 5 个子集，每次选择其中 4 个子集作为训练集，剩余 1 个子集作为验证集，重复这一过程 5 次，以保证每个子集都能作为验证集一次。最终，将 5 次验证的评估结果取平均值，作为模型的性能指标。此外，为了进一步了解模型的稳定性，我们计算了 AUC（Area Under the Curve, ROC 曲线下的面积）分数的标准差。AUC 是一种常用的二分类模型评估指标，衡量模型在不同决策阈值下区分正类和负类的能力，值越接近 1 表示模型性能越优。各模型的平均 AUC 及标准差如表 2 所示。

模型	平均 AUC	标准差
LightGBM	0.8871	0.0058
决策树	0.7931	0.0105
随机森林	0.8896	0.0053
XGBoost	0.8823	0.0056

表 2. 各模型的平均 AUC 及标准差

从平均 AUC 来看，随机森林（AUC=0.8896）和 LightGBM（AUC=0.8871）表现最优，说明这两种模型在当前数据集上能够更好地捕捉特征与目标变量之间的复杂关系，具有较强的分类能力。XGBoost 略逊一筹，AUC 为 0.8823，但依然表现出较强的预测能力。相比之下，决策树模型的平均 AUC 为 0.7931，明显低于其他三种模型，表明单一决策树的学习能力有限，难以胜任当前复杂任务。

从标准差来看，随机森林（标准差=0.0053）和 XGBoost（标准差=0.0056）的波动最小，表现出较高的稳定性，说明这些模型对数据划分的不确定性具有较强的鲁棒性。LightGBM 的标准差为 0.0058，略高于随机森林和 XGBoost，但仍处于稳定范围内。而决策树的标准差为 0.0105，波动较大，表明其性能受数据划分的影响更为显著，可靠性较低。

2.6 堆叠集成模型

堆叠集成（Stacking）是一种模型集成方法，其核心思想是通过结合多个基础学习器（Base Learners）的预测结果，输入到一个元学习器（Meta Learner）中进行二次学习，从而提升模型的泛化能力。堆叠模型的优势在于能够综合多个模型的特点，发挥模型间的互补性以提高最终分类性能。在堆叠方法中，基础学习器独立训练并生成预测结果，而元学习器通过这些预测结果作为新特征进行学习，生成最终的预测。与传统的加权平均或投票集成方法相比，堆叠集成具有更强的灵活性和优化潜力。然而，堆叠集成模型的成功依赖于基础模型的多样性以及元学习器对基础模型组合特性的学习能力。

为了验证堆叠集成的效果，我们首先选择了四个基础模型（决策树、随机森林、LightGBM 和 XGBoost）构建了第一个堆叠模型，并以逻辑回归作为元学习器。通过对训练集进行训练并在测试集上计算 AUC（ROC 曲线下面积），我们发现该堆叠模型的 AUC 为 0.6755。这一分数显著低于单一模型（如 LightGBM 的 0.8871 和随机森林的 0.8896）的表现，表明在当前任务中，堆叠模型未能有效提升分类性能。

具体而言，决策树作为单一模型表现最差（AUC=0.7931），其加入可能为堆叠模型引入了额外的噪声。此外，XGBoost 的表现略逊于 LightGBM（AUC=0.8823），且二者同为基于梯度提升的模型，可能在信息上存在较大重叠，难以为堆叠模型提供更多的多样性。基于此，我们重新选择了表现最优的两个模型——随机森林和 LightGBM，构建新的堆叠模型，以探索在减少基础模型数量的情况下是否能进一步提升性能。

在第二次堆叠实验中，我们选取了随机森林和 LightGBM 作为基础模型，并以逻辑回归作为元学习器。随机森林和 LightGBM 在 AUC 指标上表现最优且接近（分别为 0.8896 和 0.8871），并具有一定的互补性：随机森林在处理非线性和高维特征时表现较优，而 LightGBM 以其高效性和灵活性著称。堆叠模型的最终 AUC 为 0.8917，略高于单一模型随机森林（0.8896）和 LightGBM（0.8871）的表现。这表明，堆叠模型成功整合了随机森林和 LightGBM 的优点，进一步提升了分类性能。然而，AUC 的提升幅度仅为 0.46%，说明堆叠模型相

较于单一模型的增益较为有限。

2.6.1 LGBM 与堆叠模型

尽管堆叠模型的 AUC 略高于 LightGBM 和随机森林的单独表现，但考虑到实际应用中的部署成本和复杂性，最终我们选择了单一的 LightGBM 模型。

首先，堆叠模型需要训练多个基础模型和一个元学习器，这显著增加了计算成本和模型部署的复杂性。而 LightGBM 作为单一模型，其计算效率和部署便捷性明显更高，能够更快速地适应实际业务场景。其次，堆叠模型的性能提升幅度（0.46%）较小，这一增益不足以弥补其在计算资源和时间成本上的增加。最后，LightGBM 在本研究中表现稳定，其特征重要性分析提供了明确的业务解释性，对于银行制定精准营销策略具有更直接的价值。

2.6.2 LGBM 与随机森林模型

随机森林在本研究中的表现确实也略微优于 LightGBM，其平均 AUC 为 0.8896，而 LightGBM 的平均 AUC 为 0.8871。从纯粹的预测性能来看，随机森林稍胜一筹。然而，在实际应用中，选择模型不仅需要考虑性能，还需要综合考虑多个因素，包括计算效率、部署复杂性、业务需求和解释性等。

首先，LightGBM 在计算效率方面具有显著优势。LightGBM 通过基于直方的分裂算法（Histogram-based Algorithm）和叶节点增长策略（Leaf-wise Growth Strategy），能够快速处理大规模数据，同时减少内存占用。这使得 LightGBM 在模型训练和预测阶段的时间成本远低于随机森林。相比之下，随机森林需要训练大量的决策树（在本研究中为 125 棵），且每棵树的构建和预测过程均需要耗费较大的计算资源。例如，LightGBM 在本研究中的训练时间共计 8 分 54 秒。相比之下，随机森林模型的训练时间则共计 1 小时 20 分 13 秒，显著长于 LightGBM。因此在大数据环境下，随机森林的计算效率明显逊色于 LightGBM。

其次，LightGBM 提供了详细的特征重要性分析，其分裂策略能够输出每个特征对模型预测的贡献程度，为业务团队提供了直观的解释性，能够指导银行优化营销策略。相比之下，虽然随机森林也提供特征重要性，但其结果往往受到多棵树结构的复杂性影响，解释性可能不如 LightGBM 直观和细致。

最后，随机森林由于在训练过程中使用了“袋装法”（Bagging）和特征随机选择，模型的稳定性通常较好。然而，在本研究的交叉验证中，LightGBM 的表现也非常稳定，其 AU

C 的标准差为 0.0058，与随机森林的 0.0053 接近。这表明在当前数据集上，LightGBM 的表现同样可靠，其泛化能力足以胜任实际业务场景。

所以，综合考虑性能、成本和实际应用需求，我们最终选择了 LightGBM 作为预测客户是否认购银行产品的最佳模型。

3. 特征重要性分析

特征重要性分析是机器学习模型解释性研究的重要部分，其目标是量化每个特征对模型预测的贡献程度。在本研究中，我们使用 LightGBM 的内置特征重要性工具，通过 plot_importance 方法对模型的特征贡献进行可视化。这一方法基于模型在每次分裂时的增益值（Gain）来评估特征的重要性，即一个特征在决策树分裂过程中对信息增益的贡献累积值越大，其相对重要性越高。特征重要性分析不仅有助于理解模型的决策机制，还能为实际业务提供数据驱动的指导建议。在分析过程中，我们采用了 LightGBM 训练好的模型，并提取其特征贡献值，通过条形图展示出每个特征的重要性排名。

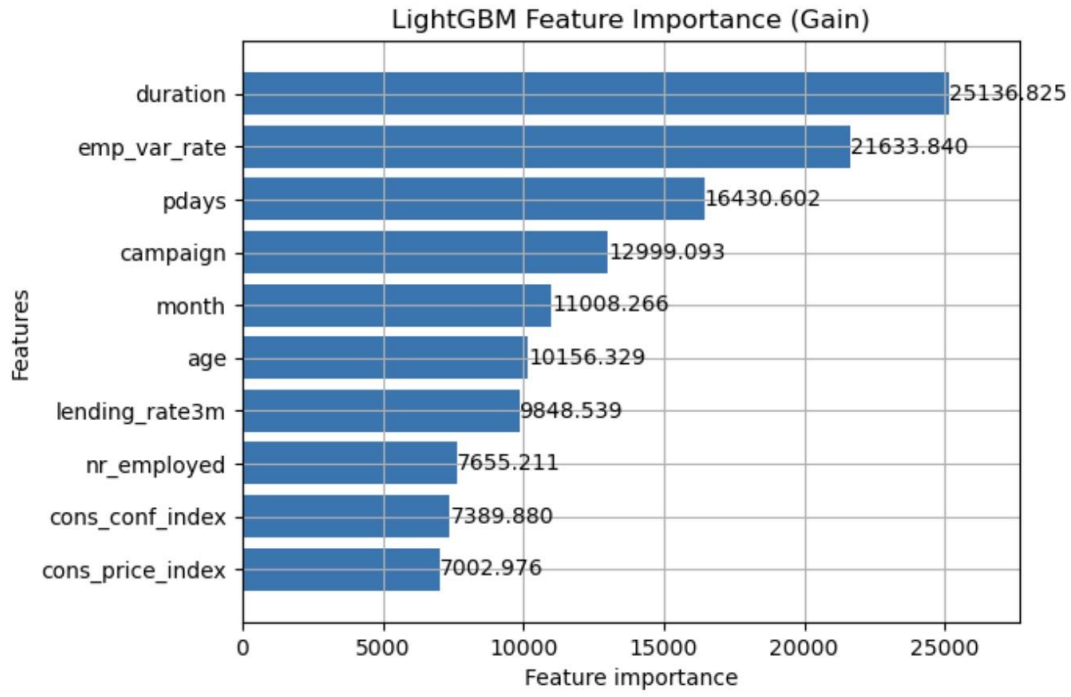

图 1. 每个特征的相对贡献

特征重要性分析的结果如图 1 所示，不同特征对客户是否购买银行产品的预测贡献存在显著差异。其中，duration（客户上一次与银行交互的持续时间）是最重要的特征，其重要性得分显著高于其他特征。这表明该特征对模型预测的影响最大。其次是 emp_var_rate（就

业变化率）和 pdays（上一次联系客户的时间间隔），这两项特征分别与宏观经济情况和客户的历史交互情况有关。此外，campaign（当前营销活动期间与客户的联系次数）和 month（上一次联系客户的月份）也表现出了较高的重要性，反映出营销活动相关的特征对预测客户行为具有重要价值。相对而言，cons_price_index（消费者价格指数）和 cons_conf_index（消费者信心指数）等宏观经济特征的贡献较低，但也具有一定的预测价值。

特征重要性分析结果揭示了客户行为的关键影响因素。首先，duration 作为最重要的特征，表明客户与银行交互的持续时间在预测客户购买意愿中起到了核心作用。这可能是因为更长的交互时间通常表明客户对产品更感兴趣或更倾向于接受银行的推荐。其次，emp_var_rate 和 pdays 的高重要性反映了宏观经济环境和客户联系历史对购买行为的重要影响。就业变化率作为经济活动的关键指标，能够间接反映客户的消费能力和购买倾向；而短时间间隔的再次联系可能激发客户的持续兴趣，进一步提升其购买可能性。此外，campaign 和 month 的重要性表明，营销活动的频率和季节性因素对客户决策具有一定的推动作用，这与银行营销实践中的经验一致。

讨论

本研究聚焦于利用机器学习技术构建和优化预测模型，评估客户是否会购买银行的产品。通过对多个主流分类模型的训练和评估，包括决策树、随机森林、LightGBM 和 XGBoost，我们发现不同模型在分类性能上存在显著差异。其中，随机森林和 LightGBM 表现优异，分别实现了 0.8896 和 0.8871 的 AUC 分数，而决策树和 XGBoost 的表现相对较差，AUC 分别为 0.7931 和 0.8823。此外，为了进一步提升模型性能，我们尝试了堆叠集成方法，将随机森林和 LightGBM 的预测结果输入到逻辑回归元学习器中。堆叠模型的 AUC 达到了 0.8917，虽然略高于单一模型的表现，但考虑到实际应用中的计算成本和部署复杂性，最终选择了性能稳定且高效的 LightGBM 作为部署模型。通过 LightGBM 的特征重要性分析，我们进一步挖掘了影响客户购买行为的关键因素，为银行营销策略优化提供了数据支持。

基于 LightGBM 的特征重要性分析结果，duration（客户与银行交互的持续时间）被确定为最重要的特征，这表明客户在交互中投入的时间越多，其购买意愿越强。银行应重视客户交互的过程，通过提高客户体验和沟通效率，延长高质量的互动时长，增加购买转化率。其次，emp_var_rate（就业变化率）和 pdays（上次联系客户的时间间隔）也对模型预测贡献较大。就业变化率反映了宏观经济环境对客户消费意愿的影响，银行可根据经济趋势调整

产品策略以适应市场变化；而联系时间间隔较短的客户往往具有更高的兴趣，银行可据此合理安排跟进周期，避免客户流失。此外，campaign（联系次数）和 month（联系月份）的重要性表明，营销活动的频率和季节性因素对客户购买行为也有显著影响。银行可以根据不同年龄段、月份以及客户特点设计定制化的营销活动，从而提高推广效率和客户满意度。

本研究在银行产品客户预测领域中取得了有价值的成果，但仍有一些不足值得进一步探索。比如，堆叠模型在本研究中表现一般，其原因可能与基础模型的多样性不足以及元学习器的优化策略有关，未来可以尝试加入其他类型的模型（如神经网络）以增强堆叠模型的效果。此外，特征选择虽然基于特征重要性进行了解释，但尚未探索添加或去除特征对模型表现的具体影响，这可能是进一步提升性能的重要方向。未来的研究可以结合更多的外部数据（如客户社会行为特征）或动态数据（如实时客户活动），以提高模型的准确性和适用性。

总的来说，本研究通过机器学习技术提供了一种高效的客户购买预测方法，揭示了影响客户行为的关键因素，为银行制定精准营销策略提供了有力支持。未来，随着数据规模的扩大和算法的改进，基于数据驱动的银行产品预测模型将有望在业务优化和客户体验提升中发挥更大的价值。

附录

项目代码及结果见：https://github.com/Greenyhua/MLUS_2024/blob/main/MPLUS_Group_9.ipynb

参考文献

Amrani, H., Lazaar, M., & Kadiri, K. E. (2018). A comparative study of decision tree ID 3 and C4.5. *International Journal of Advanced Computer Science and Applications, 9* (12), 310–318.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Bello, O. A. (2023). Machine learning algorithms for credit risk assessment: an economic and financial analysis. *International Journal of Management*, *10*(1), 109-133.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, 785–794.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., … Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems, 30,* 3146–3154.

Meade, B. (1997). Enterprise One to One: Tools for Competing in the Interactive Age. *Journal of Business and Entrepreneurship*, *9*(2), 73.

Günther, O., Rezazade Mehrizi, M. H., Huysman, M., & Feldberg, F. (2014). Debating big data: A literature review on realizing value from big data. *Journal of Strategic Information Systems, 23*(3), 191–210.