Advances in Applied Mathematics 应用数学进展, 2022, 11(9), 6426-6435
Published Online September 2022 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2022.119680

Hans汉斯

基于Stacking算法的银行定期存款产品购买行为研究

郑江怀，吕卫东*，王一朵，胡陈陈

兰州交通大学数理学院，甘肃 兰州

收稿日期：2022年8月13日；录用日期：2022年9月9日；发布日期：2022年9月16日

摘 要

研究客户的购买行为以及客户的价值成为提高银行收益与优化营销策略的主要途径。本文通过研究客户购买某商业银行定期存款产品的相关数据，预测客户是否会购买该产品。使用LabelEncoding编码方法和SMOTE算法对数据进行处理。使用RFECV和GBDT算法进行特征选择，根据特征重要性获得影响客户是否购买的重要指标。使用决策树、SVM与GBDT算法以及Stacking算法对银行客户是否会购买定期存款产品进行研究，结果显示Stacking算法的预测效果比单一模型的预测效果更好。

关键词

不平衡数据处理，Stacking算法，准确率

Purchase Behavior of Bank Time Deposit Products Based on Stacking Algorithm

Jianghuai Zheng, Weidong Lv*, Yiduo Wang, Chenchen Hu

School of Mathematics and Physics, Lanzhou Jiaotong University, Lanzhou Gansu

Received: Aug. 13th, 2022; accepted: Sep. 9th, 2022; published: Sep. 16th, 2022

Abstract

The research of customer's purchase behavior and customer's value has become the main way to improve bank's income and optimize the marketing strategy. This paper studies the relevant data of customers' purchase of fixed deposit products of a commercial bank to predict whether customers will buy this product. SMOTE algorithm and Label Encoding were used to process the data.

*通讯作者。

文章引用: 郑江怀, 吕卫东, 王一朵, 胡陈陈. 基于 Stacking 算法的银行定期存款产品购买行为研究[J]. 应用数学进展, 2022, 11(9): 6426-6435. DOI: 10.12677/aam.2022.119680

RFECV and GBDT algorithms are used for feature selection, and important indicators affecting whether customers buy are obtained according to the importance of features. Decision tree, SVM, GBDT algorithm and Stacking algorithm were used to study whether bank customers would buy time deposit products. The results showed that the prediction effect of Stacking algorithm was better than that of a single model.

Keywords

Unbalanced Data Processing, Stacking Algorithm, Accuracy

1. 引言

随着经济水平的快速提高，各企业间的竞争也越来越大。与其他企业不同，商业银行作为一种中介组织，其功能是将具有高流动性的存款转化为低流动性的信贷资产，从而为经济发展提供资本力量。商业银行是以经营存贷款业务并以此为依托全面展开金融业务的中介机构，存款是商业银行的主要资金来源[1]。银行客户则是银行存款的主要来源，于是，挖掘银行高价值客户成为了银行优化营销策略的手段之一。在 2019 年，张利利[2]等使用决策树与 BP 神经网络对银行客户进行预测，结果显示 BP 神经网络的预测效果更好。在 2021 年，王慧宇[3]通过逻辑回归与 GBDT (梯度提升树)算法对客户是否会订购银行产品进行预测，结果显示 GBDT 算法的预测效果最好。本文选择以决策树、SVM (支持向量机)与 GBDT 算法为初级分类器，以逻辑回归为次级分类器，基于 Stacking 算法对银行客户是否会购买定期存款产品进行研究。

2. 数据集

2.1. 数据集介绍

本文所使用的数据集是购买某商业银行定期存款产品的相关数据，数据来源于和鲸社区网站。银行机构采用电话呼叫电话联系人进行营销活动，在营销活动中，银行工作人员向客户推荐银行的定期存款产品，并确认客户是否会购买银行定期存款产品。客户通过参与这种活动，可以对银行的定期存款产品有更多的了解，根据自己所了解的信息选择适合自己的产品，获得一定的利息收益。对银行而言也可以提高银行的收益。该数据集共有 25317 个样本，样本中包含参与过上次活动的一些客户，其余客户均为新顾客。18 个相关指标，除了 ID 与标签以外，包含 9 个分类变量和 7 个连续性变量，标签值 y 表示客户是否购买了该产品，yes 表示购买，no 表示没有购买。如表 1 所示：

Table 1. Variable names and types
表 1. 变量名称及类型

序号	变量名	变量类型
0	Age (年龄)	连续变量
1	Job (职业)	分类变量
2	Marital (婚姻状况)	分类变量

Continued

3	Education (受教育水平)	分类变量
4	Default (是否有违约记录)	分类变量
5	Balance (每年账户的平均余额)	连续变量
6	Housing (是否有住房贷款)	分类变量
7	Loan (是否有个人贷款)	分类变量
8	Contact (联系客户的方式)	分类变量
9	Day (最后一次联系的时间 (几号))	连续变量
10	Month (最后一次联系的时间 (月份))	分类变量
11	Duration (最后一次联系的交流时常)	连续变量
12	Campaign (在本次活动中，与客户交流过的次数)	连续变量
13	Pdays (距离上次活动最后一次联系客户，过去了多久)	连续变量
14	Previous (在本次活动之前，与客户交流过的次数)	连续变量
15	Poutcome (上一次活动的结果)	分类变量

该数据集中样本的职业(Job)涉及多个行业，分别有管理层、技术人员、蓝领、学生、女佣，还有一些是失业以及退休的人。与客户联系的方式(Contact)包括移动电话和固定电话。最后一次联系的交流时长(Duration)是指在本次营销活动中与客户最后一次交流用了多长时间(单位：秒)。上一次的活动结果(Poutcome)是指在上次的营销活动中，客户是否选择购买定期存款产品。

2.2. 数据可视化分析

下面我们对数据进行可视化，对连续性变量进行描述性统计，对分类变量作图分析数据集的特点：

表 2 是关于连续性变量的描述性特征，由表中数据可知，客户的平均年龄为 40.94 岁。与客户最后一次交流时长的平均值为 258 秒，说明大多数客户对该产品是比较感兴趣的，不会拒绝银行机构的介绍。客户每年平均余额有较大的差异。其中距离上次活动最后一次联系客户，过去了多久的最小值为-1，表示与该客户在此次活动之前没有联系过，属于首次联系的客户。

由图 1 可知，在与客户进行沟通的过程中，以移动电话的方式进行沟通的客户占所有购买该产品人数的 82.44%，说明客户更容易接受这种联系方式，通过此种方式进行营销活动更容易成功。在住房贷款方面，有住房贷款与没有住房贷款的人数差异不大，其中没有住房贷款的人数占 64.24%，有住房贷款的人数占 35.76%。说明没有住房贷款的人经济比较宽裕，会考虑购买该定期存款产品。

Table 2. Characteristics of continuous variables
表 2. 连续变量特征

	age	balance	day	duration	campaign	pdays	previous
mean	40.9354	1357.5550	15.8353	257.7324	2.7721	40.2488	0.5917
std	10.6343	2999.8228	8.3195	256.9752	3.1361	100.2135	2.5683
min	18.0000	−8019.0000	1.0000	0.0000	1.0000	−1.0000	0.0000
max	95.0000	102127.0000	31.0000	3881	55.0000	854.0000	275.0000

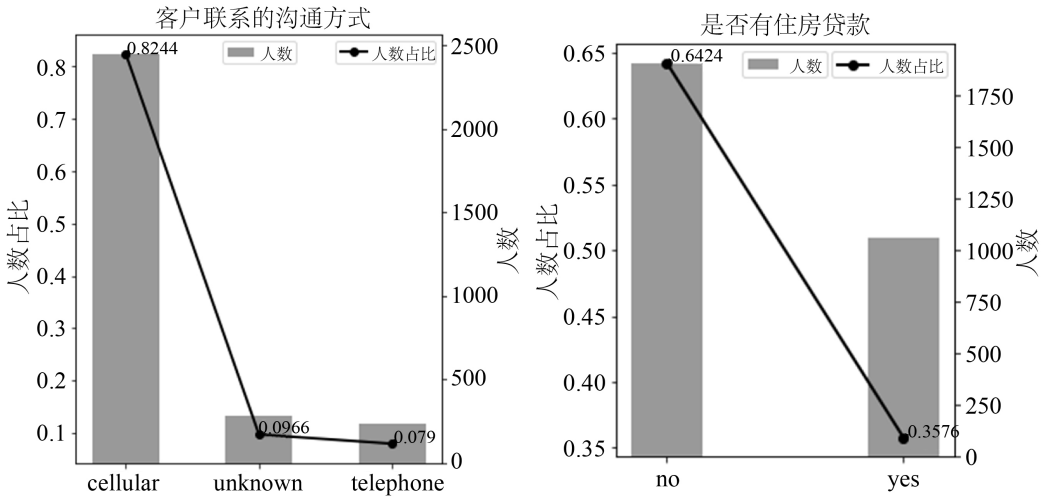

Figure 1. The communication mode of customer contact and the proportion of people who have housing loan or not
图 1. 客户联系的沟通方式和是否有住房贷款的人数占比

由图 2 可知，在此次活动选择购买该产品的客户中，有 90.54%的客户是没有个人贷款的，只有 9.46%的客户有个人贷款，说明没有个人贷款的人经济水平较高，这些客户更有可能选择购买该产品。在所有的客户中，88.3%的客户在本次活动中没有选择购买该产品，只有 11.7%的客户选择购买了该产品，说明该数据集不平衡。

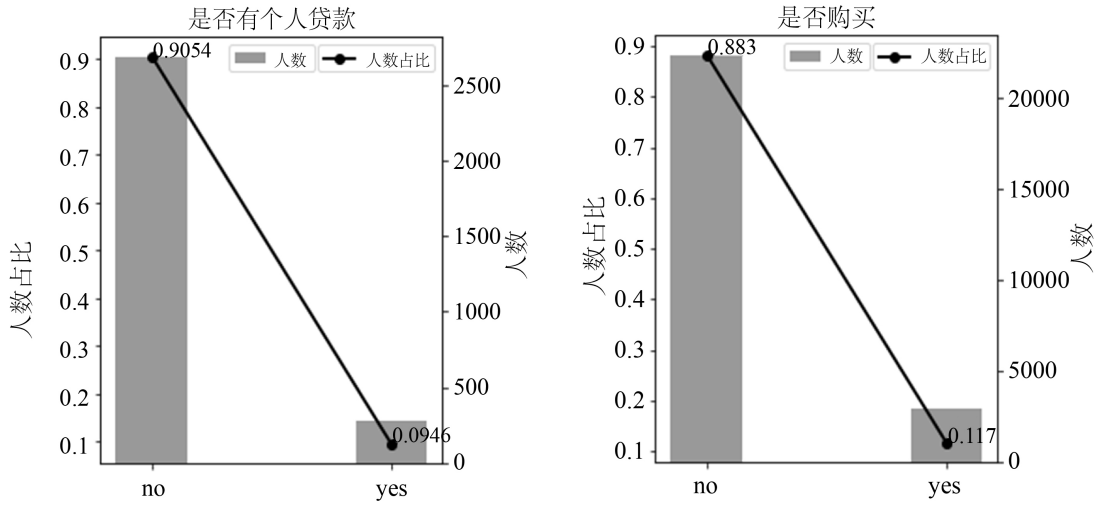

Figure 2. The proportion of whether customers have personal loans and whether they buy time deposit products
图 2. 客户是否有个人贷款与是否购买定期存款产品的人数占比

由图 3 可知，在此次活动中选择够买该产品的客户，有 18.3%的客户参与了上次的活动，并在上次活动中成功购买银行营销的产品，这说明我们可以在营销活动中联系老顾客，老顾客成功的可能性更大一些。在受教育水平方面，选择购买该产品的客户中，大部分人的受教育水平是中高等教育水平，占总人数的 83.69%，说明教育水平偏高的人更有可能选择购买该产品。

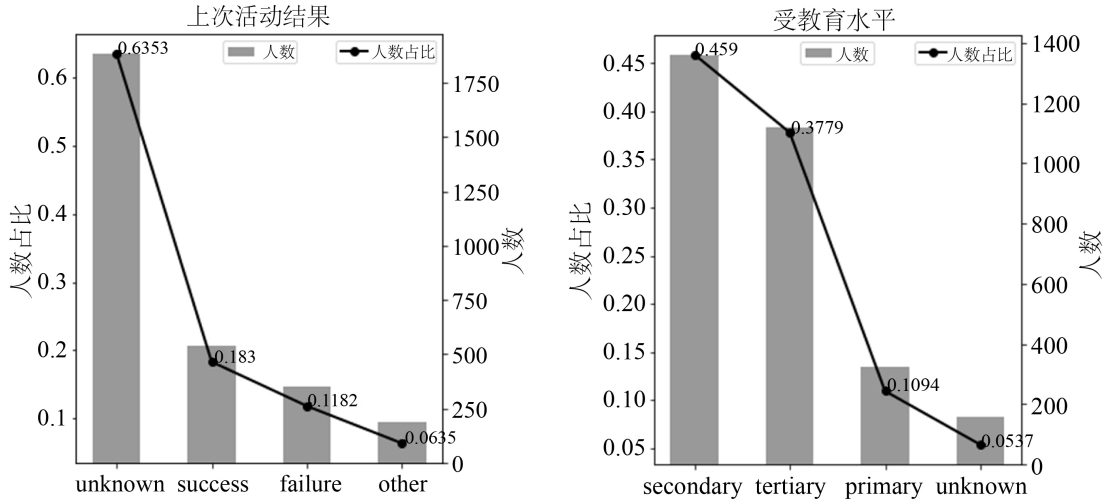

Figure 3. Ratio of the last activity result and the number of people with education level
图 3. 客户上次活动结果与受教育水平人数占比

由图 4 可知，在选择购买该产品的客户中，24.86%的客户的职业为经理，16.24%客户的职业为技术员，13.07%的客户的职业为蓝领，而服务行业的人占 10.06%，这种现象可能与职业性质有关，不同的职业收入不同，收入高的客户则经济水平更高一些，选择购买定期存款产品的可能性更大。

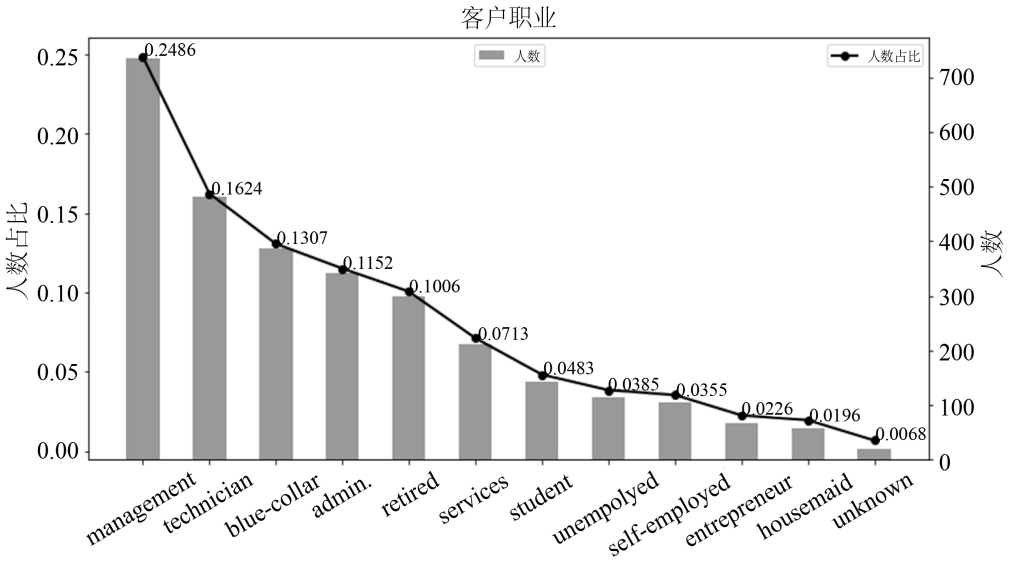

Figure 4. The proportion of different occupation types of clients
图 4. 客户不同职业类型人数占比

2.3. 数据处理与特征选择

在原始的数据集中，分类指标的取值不是数值，例如指标婚姻状况，divorced 表示离婚，married 表示已婚，single 表示单身。为了方便使用数据训练模型，我们首先需要将分类指标的取值进行数值化处理。在这里选择使用 LabelEncoding 编码方式将数据进行数值化，在该编码过程中，会为每个类别分配一个从 1 到 N 的值，N 为特征的类别数。

在实际的学术研究中，大部分分类模型都会有很好的分类效果，但是对于类别不平衡的数据，有些分类模型的分类效果会明显的变差，因此需要对数据进行平衡化处理，以提高模型的分类效果。SMOTE (Synthetic Minority Oversampling Technique)是人工合成少类样本的方法，因随机过抽样往往会导致模型过拟合且泛化能力较弱等现象，SMOTE 是对该算法的改进。它的采样原理主要是通过在一些位置比较相近的小类样本中生成新样本来达到平衡类别的目的，因为不是单方面的对小类样本进行复制，所以可以在一定程度上避免分类器的过度拟合[4]。

在数据集中，88.3%的客户没有选择购买该产品，只有 11.7%的客户选择购买该产品，由此可见，该数据集是严重不平衡的，会影响模型预测的效果，因此，在模型训练之前，需要对原始数据集进行平衡化处理，这里选择 SMOTE 算法进行平衡化处理，使得处理后的数据集中，选择购买该产品的人数与没有选择购买该产品的人数之比为 1:1，通过处理之后，得到正例样本 17,900，负例样本为 17,900。

在进行模型训练之前，需要进行特征选择。特征选择的原则是在保证模型分类精度不降低的情况下选择最少的特征。在这里，选择使用 RFECV 算法和 GBDT 算法进行特征选择。递归特征消除(recursive feature elimination, RFE)是一种搜索特征空间中最优特征子集的贪心算法，通过反复构建模型，逐次删除特征空间中模型评估重要度最低的特征以更新特征空间，直至得到所需数量的特征。交叉验证(cross-validation, CV)用于评估训练好的模型在新数据上的表现，可以在一定程度上减小过拟合，还可以从有限的数据中获取尽可能多的有效信息[5]。RFECV 在进行特征选择时主要分成两部分，一部分是递归特征消除，对特征的重要性进行评级，另一部分是交叉验证，在特征评级后，通过交叉验证，选择最佳的特征数量。如果减少特征会造成模型性能损失，将不会删除任何特征。递归特征消除从排序后的训练样本中系统的删除重要度最低的特征更新特征子集，之后对 GBDT 模型进行递归重新训练和评估。GBDT 算法属于集成学习算法的一种，它融合了装袋法(Bagging)与提升法(Boosting)的思想，由 Firedman 在 2001 年提出，既可用来解决分类问题，也可用来解决回归问题[6]。GBDT 算法主要是以迭代的方式产生多个弱分类器，每一步都在上一轮生成的分类器产生的残差上进行该阶段的训练，最终得到的分类器是对每轮训练的弱分类器输出结果进行加权处理得到最终的分类结果，该算法在数据分析和预测中有很好的效果。

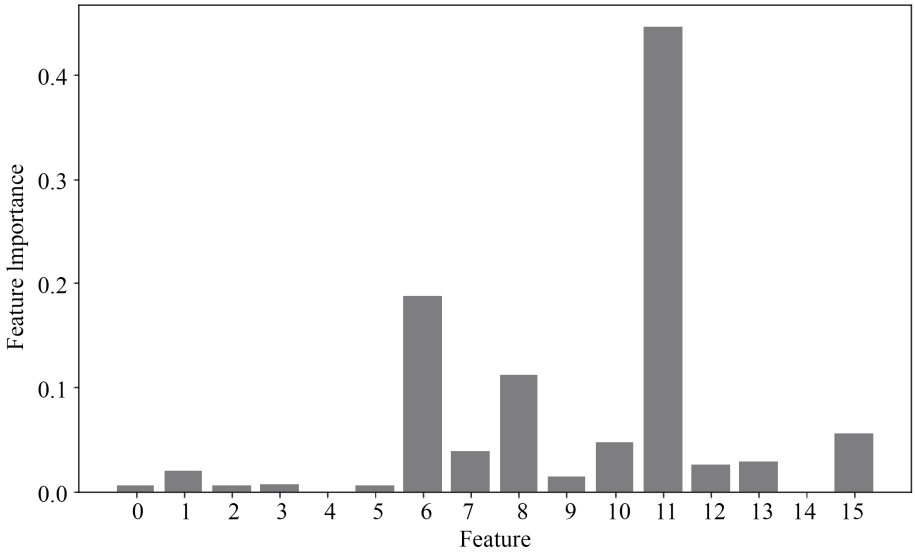

Figure 5. Feature importance of GBDT algorithm
图 5. GBDT 算法特征重要性

选择 GBDT 算法计算每个指标的特征重要性，通过重要性判断哪些指标对标签值有显著性影响。重要性值越高，则对应指标对标签值的影响越大，即该指标越重要。

在图 5 中，横坐标表示指标，其指标所对应的序号与表 1 中序号一致，纵坐标表示特征重要性的值。由图可知，最后一次联系的交流时常(Duration)的特征重要性值最大，表明该指标对标签值的影响最大。是否有住房贷款(Housing)、联系客户的方式(Contact)、上一次活动的结果(Poutcome)对标签值有较大的影响，根据之前可视化的结果可以知道，没有住房贷款的客户更有可能选择购买该定期存款产品，在与客户联系的方式方面，选择购买的客户 90%都是通过移动电话联系的。指标是否有违约记录(Default)、在本次活动之前，与客户交流过的次数(Previous)对标签值没有影响，即客户是否有违约记录不影响客户参与本次活动的结果，在本次活动之前，与客户联系次数多少也不会影响客户参与本次活动的结果。

图 6 是使用 RFECV 算法与 GBDT 算法进行特征选择的可视化结果图，横坐标表示特征个数，纵坐标表示通过交叉验证输出结果的均值，这里是五折交叉验证准确率的均值。由输出结果以及图中的曲线可知，选择 14 个特征时，五折交叉验证准确率的均值最大，因此我们选择 14 个特征训练模型，从原始特征集中剔除了特征是否有违约记录(Default)、在本次活动之前，与客户交流过的次数(Previous)。

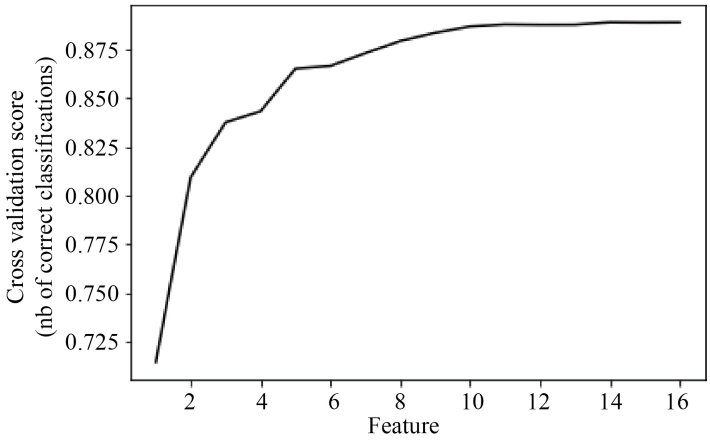

Figure 6. Feature selection of RFECV and GBDT algorithm
图 6. RFECV 与 GBDT 算法特征选择

3. 模型选择与结果分析

3.1. 模型选择

该数据集的标签是是否购买银行的定期存款产品，所以是一个二分类问题。在机器学习相关的算法中，决策树、随机森林(RF)、支持向量机(SVM)，梯度提升决策树(GBDT)等算法既可以解决二分类问题，也可以解决多分类问题，还可以解决回归类问题。这些算法在解决二分类问题时都有很好的性能。通过已有的研究发现，进行模型融合可以提高模型预测的正确率，Stacking 算法则是模型融合经典的算法之一。

Stacking 是 Wolpert 于 1992 年提出的一种集成学习的机器学习方法，它是利用原始数据集训练第 1 层的几个初级学习器，以获得多个预测结果；然后第 1 层得到的预测结果被第 2 层次级学习器作为输入特征进行训练，得到最终预测结果，从而提高模型的准确度，减少泛化误差[7]。在选择初级学习器时，一般要求选择的学习器要“好”而“不同”，“好”表示初级学习器要有较好的性能，“不同”表示初级学习器间的原理要有所差异。

因此，考虑到样本大小与特征个数，选择决策树、SVM 和 GBDT 作为 Stacking 算法的初级学习器，Stacking 算法是先训练多个初级模型，再将初级模型的输出作为次级模型的输入，一般情况下，次级模型为逻辑回归算法。不同学习算法原理不同，如逻辑回归是基于概率的分类算法，而决策树是基于树结构的分类与回归算法，融合多个算法可以多角度挖掘更多有价值的信息，降低学习算法的误差，提高预测准确率。

3.1.1. 决策树

决策树算法是分类算法中比较常用的经典算法之一，决策树算法计算复杂度不高，对中间值的缺失不敏感，得到的结果理解性强，容易看懂。建立决策树首先是分析和处理样本集中的数据，并且通过归纳算法生成一系列的规则和决策树。然后，通过使用决策树来分析新数据。决策树就是利用一组规则把数据分到不同类别中的过程[8]。

3.1.2. 逻辑回归

逻辑回归的思想是先根据边界类型构建一个边界模型和一个预测函数，再依据预测函数建立一个预测函数的损失函数，最后通过优化方法找到边界的最佳回归系数[9]。逻辑回归模型能有效地解决二分类问题，虽然是逻辑回归，但实际上是一种分类学习方法。逻辑回归可以直接对分类可能性进行预测，无需事先假设数据分布，这样就避免了假设分布不准确带来的问题。

3.1.3. SVM 算法

支持向量机是在统计学习理论的基础上，以结构风险最小化为原则建立起来的机器学习算法，通过控制参数自动调节模型结构，实现经验风险和结构风险最小化。SVM (支持向量机)是一种有监督的学习算法，主要用于解决二分类问题，对于多分类问题效果比较差，在解决中小型数据样本、非线性、高维的分类问题时具有良好的泛化能力和预测性能，在数据挖掘、图像处理、语音识别等方面应用广泛。

对于线性可分问题，支持向量机运用优化算法实现最大化分类间隔；而对于非线性问题，支持向量机通过适当的核函数将输入空间映射到高维空间，实现高维空间线性可分，将非线性问题转化线性问题，然后在新空间中利用二次型寻优算法求取最优线性分类面，从而将两类样本区分开来[10]。

3.1.4. GBDT 算法

GBDT 算法是树的一种，相对于一般的决策树算法具有防止过拟合，泛化能力较强等特点，其基本思想是通过构建 m 个弱分类器，将弱分类器线性组合为一个强分类器，每增加一个弱分类器是为了减少上一个模型的残差，并在其梯度方向上建立新的组合模型[11]。GBDT 算法在处理多特征输入分类与回归问题上表现优异，模型训练速度快，精度较高。

3.2. 结果分析

根据数据集特点选择了合适的模型之后，对数据集进行划分，数据集的 80%作为训练集，20%作为测试集。使用训练集得到模型，并用测试机测试模型的性能，模型的评价指标值如表 3 所示。

由表 3 可知，使用逻辑回归、决策树、SVM 和 GBDT 算法分别训练模型，从准确率、召回率与 F1-score 值来说，GBDT 算法训练得到的模型的性能最好。

以决策树、SVM 和 GBDT 算法为初级分类器，逻辑回归算法为次级分类器，使用 Stacking 算法进行训练模型，由表 3 可知，Stacking 算法训练得到的模型的准确率、召回率和 F1-score 比单个分来器模型的高，说明与单个分类器相比，Stacking 算法有更好的预测效果。

通过对客户购买银行定期存款产品数据的研究，不仅给客户带来方便，也给银行带来了收益。就客户而言，银行工作者在与客户交流过程中，可以根据研究获得的重要因素、模型评估客户的购买可能性，

对购买可能性很低的客户，就可以放弃继续向客户推荐银行产品，这样不会打扰到顾客，也减轻了银行工作者的工作负担，减少盲目销售带来的成本和损失。对银行而言，通过分析客户的需求，找到购买银行定期存款产品可能性较大的客户，提升客户转化率，提高银行的收益。

Table 3. Model evaluation indicators
表 3. 模型评价指标

模型	准确率	召回率	F1-score
逻辑回归	0.7984	0.80	0.83
决策树	0.8461	0.85	0.86
SVM	0.7841	0.78	0.81
GBDT 算法	0.8556	0.86	0.87
Stacking 算法	0.8736	0.87	0.88

4. 总结

随着社会的快速发展，人们在日常生活中，通过电子设备与他人进行联系，以及通过电子设备订购产品的现象越来越普遍。人们对购买产品的质量和服务也有了更高的要求，人们会经常浏览多个商品进行比较，然后选择订购自己最满意的一款产品。基于这种现象，关于客户是否购买某产品的相关数据，出现严重不平衡的现象，即大多数客户选择不购买，只有少部分客户选择购买。本文使用 SMOTE 算法对数据平衡化处理进行研究，主要结论如下：

1) 通过数据可视化分析可得，在此次活动中，与客户进行联系沟通时，以移动电话的方式进行沟通、没有个人贷款、受教育水平偏高的客户选择购买定期存款产品的可能性更大一些。在工作方面，选择购买的客户中 24.86%的客户的职业为经理，16.24%的客户的职业为技术员，13.07%的客户的职业为蓝领，这意味着工作越好，收入也越好，客户才更有可能选择购买定期存款产品。

2) 使用 Label Encoding 对分类变量进行编码后进行特征选择，根据特征重要性可知，最后一次联系的交流时常(Duration)的特征重要性值最大，表明该指标对标签值的影响最大。是否有住房贷款(Housing)、联系客户的方式(Contact)、上一次活动的结果(Poutcome)对标签值有较大的影响，指标是否有违约记录(Default)，在本次活动之前，与客户交流过的次数(Previous)对标签值没有影响，即客户是否有违约记录不影响客户参与本次活动的结果，在本次活动之前，与客户联系次数多少也不会影响客户参与本次活动的结果。

3) 通过特征选择结果可知，选择 14 个指标进行训练模型，通过使用决策树、SVM 与 GBDT 算法训练得到模型，使用 Stacking 算法将单一模型融合后的预测效果比单一模型的预测效果更好，准确率为 0.8736，召回率为 0.87，F1-score 的值为 0.88。

致　谢

本文章的顺利完成，感谢国家自然科学基金的项目支持，感谢我的导师吕卫东副教授，谢谢他对我的悉心指导，他无私的关爱和严谨的治学态度将激励我不断进取。还要感谢我的朋友王一朵和胡陈陈的帮助与启迪，是他们的激励和支持，让我对完成这篇文章更有信心。

基金项目

国家自然科学基金项目(11961039)。

参考文献

[1] 郭铭文. 商业银行存款稳定性研究[D]: [博士学位论文]. 吉林: 吉林大学, 2006.

[2] 张利利, 郭淑妹, 马艳琴, 卜春霞. 基于数据挖掘技术的银行客户定期存款认购模型研究[J]. 数学的实践与认识, 2019, 49(21): 95-102.

[3] 秦胜伍, 张延庆, 张领帅, 苗强, 程秋实, 苏刚, 孙镜博. 基于Stacking模型融合的深基坑地面沉降预测[J]. 吉林大学学报(地球科学版), 2021, 51(5): 1316-1323.

[4] 张鹏. 面向不平衡数据的分类技术研究及应用[D]: [硕士学位论文]. 太原: 山西财经大学, 2021.

[5] 常家康, 吕宁, 詹跃东. 基于XGBoost-RFECV算法和LSTM神经网络的PEMFC剩余寿命预测[J]. 电子测量与仪器学报, 2022, 36(1): 126-133.

[6] 周长春, 姜杰, 李谦, 朱海燕, 李之军, 鲁柳利. 基于融合特征选择算法的钻速预测模型研究[J]. 钻探工程, 2022, 49(4): 31-40.

[7] 王慧宇. 基于机器学习预测市场营销活动对客户订购定期存款的影响[D]: [硕士学位论文]. 天津: 南开大学, 2021.

[8] 刘云翔, 吴浩. 基于改进CART决策树建立水华预警模型[J]. 中国农村水利水电, 2018(1): 26-28.

[9] 庞文琦. 基于逻辑回归的内齿轮测量[D]: [硕士学位论文]. 沈阳: 沈阳工业大学, 2021.

[10] 张松兰. 支持向量机的算法及应用综述[J]. 江苏理工学院学报, 2016, 22(2): 14-17+21.

[11] 徐永瑞. 基于机器学习算法的用电量预测[D]: [硕士学位论文]. 天津: 天津大学, 2020.