Statistics and Application 统计学与应用, 2025, 14(6), 383-395
Published Online June 2025 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2025.146176

Hans汉斯

基于银行客户订购金融产品指标的多种机器学习预测模型构建及比较研究

詹长超

广西师范大学数学与统计学院，广西 桂林

收稿日期：2025年5月30日；录用日期：2025年6月22日；发布日期：2025年6月30日

摘　要

本研究旨在通过分析银行客户的多种指标，预测其是否会订购银行金融产品。针对这一分类问题，采用了多种机器学习方法，包括Logistic回归模型、支持向量机(SVM)模型和随机森林算法，对客户数据进行了深入分析。首先，研究对银行客户数据进行预处理，包括数据清洗、缺失值处理、变量分类、变量选择等，以确保数据质量。其次，采用Logistic回归模型、支持向量机(SVM)、随机森林模型在数据集上进行训练和测试。最后将Logistic回归模型、支持向量机(SVM)、随机森林模型根据训练和测试的结果对它们的性能效果进行对比分析。通过比较混淆矩阵，ROC曲线，预测精确度等，确定了最优模型。结果表明：随机森林在银行客户订购金融产品的分类问题上具有较强的预测能力。年龄、职业、婚姻状况、教育水平等因素对客户是否订购金融产品有显著影响。

关键词

金融产品，机器学习，Logistics回归，比较研究

Construction and Comparative Study of Multiple Machine Learning Prediction Models Based on Bank Customer Metrics for Financial Product Subscription

Changchao Zhan

School of Mathematics and Statistics, Guangxi Normal University, Guilin Guangxi

Received: May 30th, 2025; accepted: Jun. 22nd, 2025; published: Jun. 30th, 2025

文章引用: 詹长超. 基于银行客户订购金融产品指标的多种机器学习预测模型构建及比较研究[J]. 统计学与应用, 2025, 14(6): 383-395. DOI: 10.12677/sa.2025.146176

Abstract

This study aims to predict whether bank customers will subscribe to financial products by analyzing multiple customer metrics. To address this classification problem, various machine learning methods—including logistic regression, support vector machines (SVM), and random forest algorithms—were applied for in-depth data analysis. First, the bank customer data underwent preprocessing, including data cleaning, missing value imputation, variable categorization, and feature selection, to ensure data quality. Subsequently, logistic regression, SVM, and random forest models were trained and tested on the dataset. Finally, the performance of these models was compared and analyzed based on training and testing results. By evaluating confusion matrices, ROC curves, prediction accuracy, and other metrics, the optimal model was identified. The results indicate that the random forest algorithm demonstrates strong predictive capability for classifying bank customers' financial product subscriptions. Factors such as age, occupation, marital status, and education level significantly influence customers' subscription decisions.

Keywords

Financial Products, Machine Learning, Logistic Regression, Comparative Study

1. 引言

在高度竞争的金融服务市场中，银行和其他金融机构正寻求新方法来提高其营销策略的效率和有效性。理解客户的需求和行为，并据此预测他们是否会订购某个金融产品，一直是银行营销管理的核心挑战。传统的营销方法可能会消耗大量资源，但转化率往往不高，导致营销成本居高不下。随着大数据和人工智能技术的发展，机器学习已经成为提高预测准确度、优化资源分配、并量化决策支持的有力工具。特别是在金融领域，利用客户的历史数据和交易行为来构建预测模型，可以在一定程度上帮助银行识别最有可能订购新金融产品的客群，从而针对性地进行市场营销，提高营销的 ROI (返回投资比率)。此外，监管要求的透明度和合规性也对银行的业务操作和客户关系管理提出了新的要求。在这种背景下，研究和比较不同机器学习模型的预测效果，对于银行来说至关重要，可以帮助它们更好地分析风险和机会、作出基于数据的决策，并提高整体的客户服务水平[1]。

本文针对银行客户的各种指标进行分析，利用机器学习方法来预测其是否会订购金融产品。银行客户的订购行为受多种因素影响，包括但不限于个人收入水平、家庭背景、职业身份、历史购买行为以及银行产品推广策略等。因此，通过综合分析客户的多种指标来预测其订购金融产品的可能性，对于银行机构而言具有重要的实践意义。

2. 模型构建

本文以年龄、职业、客户婚姻状态、住房贷款、个人贷款等指标作为自变量，是否订购该服务作为因变量，采用 Logistics 回归构建回归模型，训练支持向量机、随机森林这两种机器学习方法，再结合三种模型结果比较哪种模型的预测效果更好。

(一) Logistic 回归

Logistic 回归模型作为一种广泛应用于分类问题的统计方法，在银行客户订购金融产品意愿的预测中扮演着重要角色。

设有一个因变量 Y 和 m 个自变量 $X_1, X_2, \cdots, X_m$，因变量 Y 是个二值变量时，取值为

$$Y = \begin{cases} 1 & \text{订购金融产品} \\ 0 & \text{不订购金融产品} \end{cases}$$

在一组自变量作用下是否订购金融产品的发生概率为 $P(Y=1 \mid X_1, X_2, \cdots, X_m)$，记为 P，则 Logistic 回归模型又可以表示为[2]

$$P = \frac{1}{1+\exp\left[-\left(\alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_m X_m\right)\right]}$$

其中 α 为常数项，$\beta_1, \beta_2, \cdots, \beta_m$ 模型参数，经过变换可以得到模型的线性化表示：

$$\ln\left(\frac{p}{1-p}\right) = \alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_m X_m$$

模型左边为订购发生的概率与不订购发生概率的比值再取自然对数，记作 logistic(P)，于是表达为[3]

$$\log it(p) = \alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_m X_m$$

记 $Z = \alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_m X_m$，那么 Z 和 P 之间关系的 logistic 曲线如下图 1。

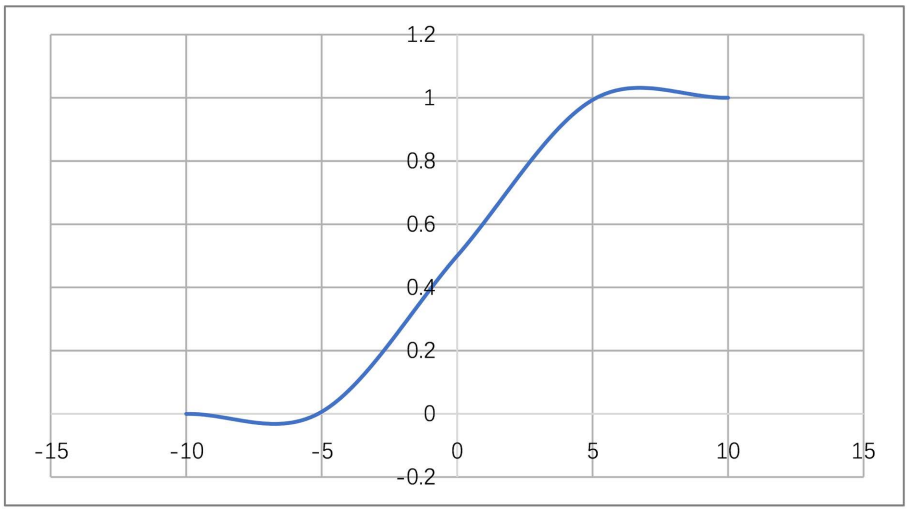

Figure 1. Logistic curve
图 1. Logistic 曲线

当 $Z \to +\infty$ 时，P 值渐近于 1，当 $Z \to -\infty$ 时，P 值渐近于 0；P 值在 0~1 的范围内变化，随 Z 值的增加或减小以点(0, 0.5)为中心呈 S 形变化。该特点使得 Logistic 模型可以较好的反映银行客户订购金融产品的趋势。

相对于其他更复杂的模型(如支持向量机或神经网络)，Logistic 回归模型的计算开销较小，训练速度快。

模型的系数可以解释特征对结果的影响程度，有助于理解数据。

(二) SVM

1995 年 Vapnik 等人在统计学习理论 SLT 的基础上首次提出了一种新的有监督机器学习方法，即支持

向量机，支持向量机(Support Vector Machine, SVM)是一种强大的监督学习算法，常用于分类和回归分析。

对于非线性问题，我们可以通过核技巧的方法来解决。先把训练集打上类别标签，让 SVM 模型用训练集进行训练，训练之后，给训练好的 SVM 模型输入测试集数据进行测试训练效果，SVM 模型就可以自动对测试集数据进行分类。

假设现在我们在学习一个非线性分类规则，首先，给出一组非线性带标签的输入样本 $\vec{x}_i$，$i=1,\cdots n$，及其对应的期望输出 $y_i \in \{+1,-1\}$，$\vec{x}_i$ 对应于转换数据点的线性分类规则 $\varphi(\vec{x}_i)$。然后，给出一个核函数 k，满足 $k(\vec{x}_i,\vec{y}_j)=\varphi(\vec{x}_i)\cdot\varphi(\vec{x}_j)$。我们找到向量 $\vec{w}$ 的分类在转换空间中满足

$$\vec{w}=\sum_{i=1}^{n} c_i y_i \varphi(\vec{x}_i)$$

其中 c_i 可以通过优化问题求解获得。对于所有 i，在约束条件 $\sum_{i=1}^{n} c_i y_i = 0$ 和 $0 \le c_i \le \frac{1}{2n\lambda}$ 下[4]，

$$f(c_1\cdots c_n)=\sum_{i=1}^{n} c_i - \frac{1}{2}\sum_{i=1}^{n} y_i c_i \left(\varphi(\vec{x}_i)\cdot\varphi(\vec{x}_j)\right) y_j c_j$$

$$=\sum_{i=1}^{n} c_i - \frac{1}{2}\sum_{i=1}^{n} y_i c_i k(\vec{x}_i,\vec{x}_j) y_j c_j$$

达到最大化。其中系数 c_i 可以通过二次规划求解获得。同样，我们可以找到一些 i 值，使得 $0<c_i<\frac{1}{2n\lambda}$，从而使得 $\varphi(\vec{x}_i)$ 位于转换空间的间隔的边界上，然后用 i 值求解

$$b=\vec{w}\cdot\varphi(\vec{x}_i)-y_i=\left[\sum_{j=1}^{n} c_j y_j \varphi(\vec{x}_j)\cdot\varphi(\vec{x}_i)\right]-y_i$$

$$=\left[\sum_{j=1}^{n} c_j y_j k(\vec{x}_j,\vec{x}_i)\right]-y_i$$

最后，新的没有带标签的非线性数据点通过此模型计算，进而达到二元线性分类目的。

$$\vec{z}\mapsto \operatorname{sgn}(\vec{w}\cdot\varphi(\vec{z})-b)=\operatorname{sgn}\left(\left[\sum_{i=1}^{n} c_i y_i k(\vec{x}_i,\vec{z})\right]-b\right)$$

(三) 随机森林

随机森林是由 Leo Breiman 和 Adele Cutler 于 2001 年提出的一种集成学习(Ensemble Learning)算法，在大数据分析、生物信息学、金融风控等领域被广泛应用。随机森林算法的核心思想是“集成学习”，当一个样本需要被分类时，它的结果不是仅仅取决于某一棵决策树得出的结果，而是让森林中的每一棵决策树都得出一个结果，再根据哪一类结果的数量最多，就将其作为最终的结果输出。随机森林算法的优点在于它对数据集的扰动具有强大的适应性，在许多方式中对不良事实具有较高的鲁棒性。此外，它也是一种可以处理高维特征的算法，因为随机森林算法可以自适应地减少对不重要特征的依赖[5]。

(四) 模型评价指标

1. 混淆矩阵(Confusion Matrix)

混淆矩阵是用来评价模型分类能力的矩阵(表 1)，一般来说是一个 $N\times N$ 的矩阵，由于本文中的因变量是一个二分类的变量，所以在本文当中我们会采用 2×2 的混淆矩阵来评估上述四种模型的分类能力。

Table 1. Confusion matrix
表 1. 混淆矩阵

	预测为负	预测为正
实际为负	TN	FP
实际为正	FN	TP

混淆矩阵里面包含了四类信息，分别是 TP (True Positive)：表明实际上是正预测也为正的样本数，真实为 0，预测也为 0。FN (False Negative)：表明实际上是负却预测为正的样本数，真实为 0，预测为 1。FP (False Positive)：表示实际上是正却预测为负的样本数，真实为 1，预测为 0。TN (True Negative)：表明实际上为负预测也为负的样本数，真实为 1，预测也为 1。

对于一个混淆矩阵仅仅只看这四类信息是完全不够的，这四类信息只是反应了一些数量，不能直接判断模型预测效果的好坏，因此还需要计算一些由混淆矩阵衍生出来评价指标，分别是：

$$\text{准确率} = \frac{TP+TN}{TP+FN+FP+TN} \qquad \text{精确度} = \frac{TP}{TP+FP}$$

$$\text{召回率} = \frac{TP}{TP+FN} \qquad \text{误报率} = \frac{FP}{FP+TN}$$

混淆矩阵在对模型预测效果的评价中最常用到的就是准确率(Accuracy, ACC)，这是指被正确预测的样本数在总样本数中所占的比重，也就是模型正确分类的概率。

2. 接受者操作特征曲线(Receiver Operating Characteristic Curve, ROC)

ROC 曲线是以召回率为纵坐标，误报率为横坐标绘制的曲线。ROC 曲线能够简单直观地反映出模型的预测效果，ROC 曲线越靠近左上角，说明模型的预测效果越好。本文将 Logistic 回归模型、支持向量机(SVM)、随机森林的 ROC 曲线绘制在同一张图中，方便能够直观地判断各个模型的预测效果。

AUC (Area Under ROC Curve)：是另一个评价二分类模型的指标，它表示的是 ROC 曲线和横轴所围成的面积。AUC > 0.5 表示模型有效果，AUC 的值越大表示模型的分类效果越好，AUC = 0.5 则说明模型无效，AUC < 0.5 则说明出现标签错误的情况。

3. 数据来源及处理

本文的数据采用的是阿里云天池数据集网站中公开的银行客户认购产品预测数据集。因变量按照是否订阅分为订阅该产品和不订阅该产品。分析指标包括名义变量：年龄、职业、婚姻状态、受教育水平、是否有信用违约记录等，数值变量：年龄、通话持续时间、本次活动联系的次数等 21 个指标变量。剔除有缺失值的观察单位，并按照表 2 的方式对名义变量进行数值替换。本文在使用同一组数据，将数据集随机地划分为数据量满足 3:1 的训练集和测试集。

Table 2. Code-variable reference table
表 2. 编码变量对照表

变量	编码	变量	编码	变量	编码
16~23 岁	0	技术员	10	无个人贷款	1
24~40 岁	1	失业者	11	个人贷款未知	2
41~60 岁	2	未知职业	12	有个人贷款	3
61 岁以上	3	基础教育，4 年制	1	使用座机	1
离婚	1	基础教育，6 年制	2	使用手机	2
结婚	2	基础教育，9 年制	3	上次营销失败	1
单身	3	高中毕业	4	没有被营销	2
婚姻未知	4	文盲	5	上次营销成功	3
行政人员	1	专业课程	6	通话时间 0~100 分钟	0
蓝领工人	2	大学学位	7	通话时间 101~200 分钟	1
企业家	3	学历未知	8	通话时间 201~500 分钟	2

续表

家政妇	4	无信用违约	1	通话时间 501~1000 分钟	3
管理层	5	信用违约未知	2	通话时间 1001~2000 分钟	4
退休人士	6	信用违约未知	3	通话时间 2001~3500 分钟	6
自由职业	7	无住房贷款	1	通话时间 3501 分钟以上	7
服务行业	8	住房贷款未知	2	不订购	0
学生	9	有住房贷款	3	订购	1

4. 结果分析

本文调用 R x64 4.3.3 中的开源包如 e1071 包、randomForest 包等构建包括 Logistics 回归、随机森林、支持向量机在内的回归模型或机器学习模型。本文通过分析模型测试集混淆矩阵计算准确率、ROC 曲线等指标综合分析、阐述模型性能，给出最终建议。

(一) 回归模型结果分析

1. Logistic 回归

本文在这一节使用 lm 函数构建 logistic 回归模型，本文共分析了 30,000 名银行客户的信息，分析各影响因素对银行客户是否订购金融产品进行分析，并在测试数据集上生成预测，根据 0.5 的阈值将预测的概率转换为二元类别，为进一步分析将预测的类别转换为具有 0 和 1 水平的因子。

Table 3. Logistic regression coefficient results
表 3. Logistics 回归生成结果系数表

	Estimate	Std. Error	z value	Pr (>|z|)
(Intercept)	−3.45	1.110000	−3.108	0.001881
id	−0.00006455	0.000003	−22.633	<2e−16
age	0.2048	0.035650	5.745	9.20E−09
job	0.03115	0.006446	4.832	1.35E−06
marital	0.2353	0.035770	6.578	4.76E−11
education	0.04676	0.011270	4.151	3.31E−05
default	0.07464	0.052920	1.41	0.158423
housing	−0.0424	0.024100	−1.76	0.078485
loan	0.05134	0.030710	1.671	0.094623
contact	−0.3051	0.060160	−5.072	3.94E−07
month	0.05096	0.010150	5.021	5.13E−07
day_of_week	0.03246	0.016460	1.973	0.048517
duration	0.1774	0.014670	12.093	<2e−16
campaign	0.04405	0.002605	16.913	<2e−16
pdays	−0.0003072	0.000081	−3.772	0.000162
previous	−0.1162	0.015220	−7.637	2.22E−14
poutcome	0.1781	0.036320	4.905	9.33E−07
emp_var_rate	−0.4413	0.018180	−24.278	<2e−16
cons_price_index	0.01877	0.008358	2.246	0.024732
cons_conf_index	0.004575	0.003735	1.225	0.220653
lending_rate3m	−0.05641	0.018590	−3.034	0.002413
nr_employed	−0.0002884	0.000145	−1.995	0.04607

基于以上 Logistics 回归模型生成的结果可知(表 3)，年龄的系数为 0.2048，表示年龄每增加一个单位(一岁)，对数几率增加 0.2048，即年龄对订阅倾向有正影响，即年龄越大客户越倾向于订购金融产品。Job、marital、education 这些分类变量的 P 值较小表明职业、婚姻状况和教育水平对于预测是否订购有显著影响。default 表示是否信用违约，其系数为 0.07464，说明有系数违约的客户更偏向于订购金融产品，housing 表示是否有住房贷款，其系数为−0.0424，表示有住房贷款的客户并不愿意订购金融产品，loan 表示是否有其他贷款，其系数为 0.05134，表示有其他贷款的客户希望订购金融产品，但是 default、housing 和 loan 三个变量的 P 值较大，表明它们对预测订购的能力较弱。Month、day_of_week 分别表示交流的时间点，月份和星期几的系数表明时间因素对顾客的订阅行为有一定影响，且他们的系数都是正的，表明月份越大银行客户成功几率越高，周一到周五成功几率递增。duration 表示通话持续时间通话持续时间的系数为 0.1774，是显著的正系数($p < 0.001$)，表明通话时间越长，顾客订阅产品的可能性越大。Campaign 表示本次活动中与客户联系的次数，Campaign 的系数为 0.04405 表示联系的次数越多，客户订购的几率越大。Pdays 表示自上次活动后，经过的天数，Pdays 的系数为−0.0003072，表示经过的天数越多，客户越不想订购金融产品，所以银行得多多宣传。previous 在本次活动之前，与该客户的联系次数，其系数为−0.1162，表示活动前和银行客户联系太多会起负作用。Poutcome 表示上一次营销活动的结果，其系数为 0.1781，表示上一次活动没成功的客户，这一次更偏向订购金融产品。Duration、Campaign、Pdays、previous、Poutcome 均显著影响订阅意愿。emp_var_rate 表示就业变动率(经济指标之一)，其参数为−0.4413，表明可能人们经常更换工作，可能面临不稳定的收入。不确定的收入使得人们更难作出长期的财务承诺，如定期支付金融产品的费用。cons_price_index 为消费者价格指数(衡量通胀或物价水平变化的指标)其参数为 0.01877，表明随着生活成本的上涨，人们可能更加关注财务规划和储蓄。他们可能会寻求购买金融产品或其他投资工具，以确保在未来能够应对可能的财务挑战。cons_conf_index 表示消费者信心指数(衡量消费者对经济状况的信心水平)，其系数为 0.004575，表明经济状况越好，客户越偏向购买金融产品。lending_rate3m 表明三个月贷款利率(可能影响贷款的吸引力)，其参数为−0.05641，表明三个月贷款利率越高，人们越不想购买金融产品。nr_employed 表示受雇人数(经济指标之一)，其参数为−0.0002884，表示当“受雇人数”增加时，与之相关的金融产品被订购的可能会减少。这样的关系可能是因为受雇人数增加意味着劳动力市场供给增加，从而导致工资水平下降或失业率上升，最终影响到了金融产品。均显著影响订阅意愿。emp_var_rate、cons_price_index、cons_conf_index、lending_rate3m、nr_employed：这些经济指标的系数表明经济环境的不同方面对顾客的订阅行为有显著影响。

接着我们分析模型的拟合质量，通过 Null deviance 和 Residual deviance 的比较显示，加入解释变量后模型的偏差显著减少，这表明模型能有效地解释数据。

Table 4. Confusion matrix for Logistic regression
表 4. Logistic 回归的混淆矩阵

	预测为 0	预测为 1
实际为 0	5391	518
实际为 1	50	41

表 4 生成的是 Logistic 模型的混淆矩阵，下面是对这个混淆矩阵相关内容的分析：

混淆矩阵中的值 5391 表示模型正确预测了 5391 个非订购的案例，值 518 表示有 27 个非订购的客户被错误预测为订购了该服务，值 50 表示有 50 个实际上订购了该服务的客户被错误地预测为没有订购该服务，值 41 表示模型正确预测的 41 个订购了该服务的案例。

从混淆矩阵包含的四部分信息能够计算出由混淆矩阵衍生出的准确率、敏感度、特异度、正例命中

率、负例命中率，数值如下：

准确率：(TP +TN)/总数 = (5391 + 41)/(5391 + 518 + 50 + 41) = 0.905

召回率：TP/(TP + FN) = 41/(41 + 50) = 0.451

误报率：FP/(FP + TN) = 518/(518 + 5391) = 0.087

精确度：TP/(TP + FP) = 41/(41 + 518) = 0.073

从上述结果来看，虽然 Logistic 模型在整体准确率上表现良好，但在误报率和精确度方面表现严重不足。表明分类器在将负例错误地分类为正例时表现不佳，导致了误报率的上升，同时降低了精确度。虽然分类器在总体上对于正负样本的分类准确率较高(准确率高)，但它在区分负例时存在较大的困难，因此在特定场景下可能无法有效地减少误报[6]。

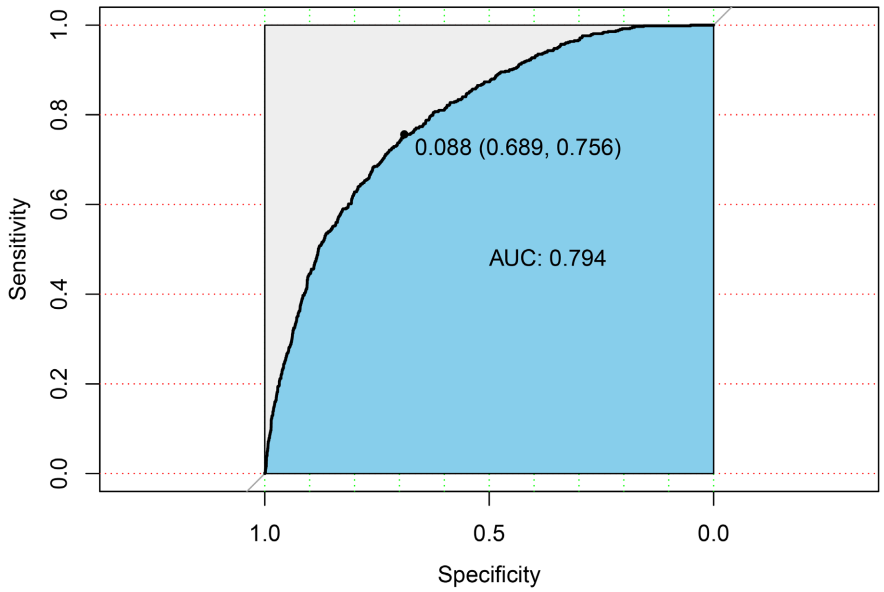

Figure 2. ROC curve of the Logistic model
图 2. Logistic 模型的 ROC 曲线

图 2 是 Logistic 模型根据训练集和测试集生成的对应 ROC 曲线。AUC 值为 0.794，表示模型的整体分类能力良好，显著高于随机分类的基准(0.5)。AUC 值接近 0.8，通常表明一个较好的分类器能够捕捉大部分区分样本是否订购该服务的信号，但还有少部分的信号未能捕捉到。在灵敏度约为 0.689 和特异度约为 0.756 的点上，标注了一个特定的分类阈值(0.088)，表明在该阈值下分类的性能表现。总体上，这张 ROC 曲线图表明了逻辑回归模型在区分正类和负类方面具有一定的可靠性，但仍有一定的提升空间。

(二) 多种机器学习方法预测结果及分析

1. 支持向量机(SVM)

本文利用 R x64 4.3.3 版本开源 e1071 进行模型建立得到以下混淆矩阵和 ROC 曲线。

Table 5. Confusion matrix for SVM
表 5. SVM 混淆矩阵

	预测为 0	预测为 1
实际为 0	6756	27
实际为 1	664	53

表 5 显示的是 SVM 模型的混淆矩阵，基于前文设定，混淆矩阵中 0 表示客户没有订购该产品，1 表示客户订购了该产品。

下面是对这个混淆矩阵相关内容的分析[7]：

混淆矩阵中的值 6756 表示模型正确预测了 6756 个非订购的案例，值 27 表示有 27 个非订购的客户被错误预测为订购了该服务，值 664 表示有 664 个实际上订购了该服务的客户被错误地预测为没有订购该服务，值 53 表示模型正确预测的 53 个订购了该服务的案例。

根据混淆矩阵包含的四部分信息能够计算出由混淆矩阵衍生出的准确率、敏感度、特异度、正例命中率、负例命中率，数值如下：

准确率：(TP + TN)/总数 = (53 + 6756)/(6756 + 27 + 664 + 53) = 0.9079

召回率：TP/(TP + FN) = 53/(53 + 664) = 0.073

误报率：FP/(FP + TN) = 27/(27 + 6756) = 0.0039

精确度：TP/(TP + FP) = 53/(53 + 27) = 0.6625

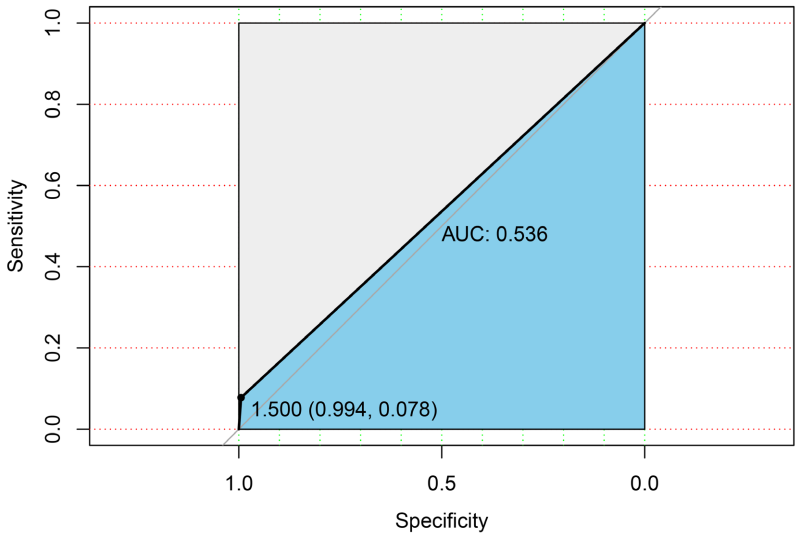

Figure 3. ROC curve of the SVM model
图 3. SVM 模型 ROC 曲线

图 3 表示的是 SVM 模型的 ROC 曲线，图中显示 SVM 模型基于测试集得出的 AUC 值为 0.536，这个值只是略大于 0.5，意味着模型的预测能力几乎和随机猜测没有区别，分类能力很不理想[8]-[10]。

从 ROC 曲线上来看这个 ROC 曲线非常接近对角线，意味着模型没有很好地区分两个类别。图中标记了一个特定阈值下的点(0.994, 0.078)，以及这个阈值对应的标准化距离(1.500)。这个点的灵敏度很低，特异性很高，意味着在这个阈值下，模型倾向于讲大多数的样本分类为未订购类，从而避免了预测错误，单页错过了许多真正订购的客户[11]。

从上述结果来看，虽然 SVM 模型在整体准确率上表现良好，但在敏感度方面表现严重不足。这表明该模型虽然较少地产生误报，但却容易错过真正的订购客户。从 SVM 模型的 ROC 曲线和 AUC 值上能看出，SVM 模型在区分订购与非订购用户方面的表现不佳[12]。

2. 随机森林

本文利用 R x64 4.3.3 版本开源 randomForest 包进行模型建立得到决策树数量和 OOB 误差关系图、混淆矩阵和 ROC 曲线。

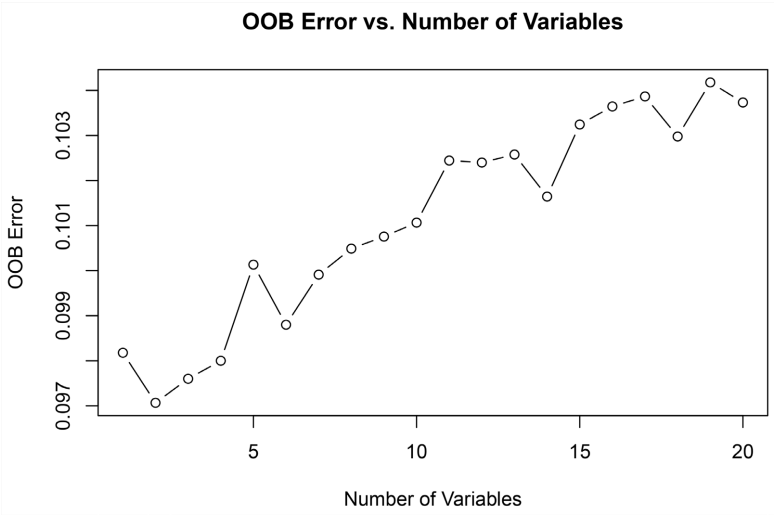

Figure 4. Relationship between node variables and OOB error
图 4. 节点变量与 OOB 误差关系图

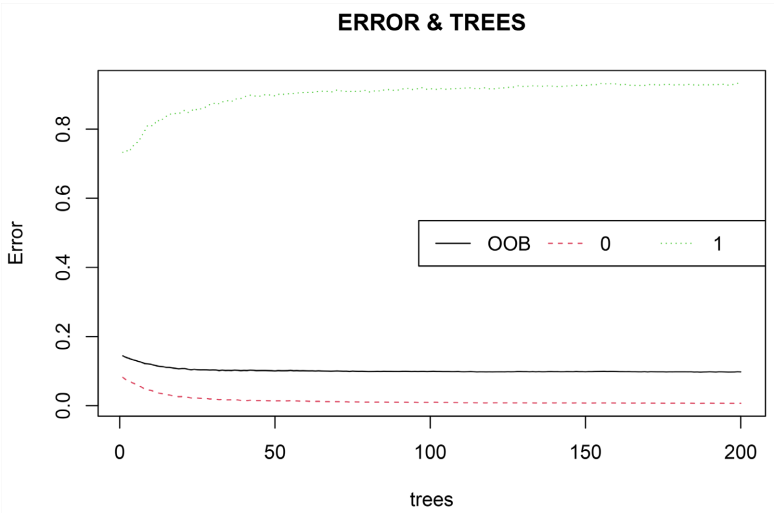

Figure 5. Relationship between the number of decision trees and OOB error
图 5. 决策树数量和 OOB 误差关系图

本文所采用的数据中含有较多的自变量，为了能够获得最佳的自变量组合，我们计算了每个节点可供选择的变量数目从 1 到 21 与误差的关系图，从图 4 中可以看出，当自变量数量取值为 2 的时候误差最低，效果最好。后续固定变量抽取的数量为 2，依次计算决策树数量从 1 到 200 的误差图 5，从图中可以看到生成 200 棵决策树时模型的误判的概率逐渐降低且趋于稳定[13]。

Table 6. Random forest confusion matrix
表 6. 随机森林混淆矩阵

	预测为 0	预测为 1
实际为 0	6548	214
实际为 1	573	165

表 6 显示的是随机森林生成的混淆矩阵，下面是对这个混淆矩阵相关内容的分析：

混淆矩阵中的值 6548 表示模型正确预测了 6548 个非订购的案例，值 214 表示有 27 个非订购的客户被错误预测为订购了该服务，值 573 表示有 573 个实际上订购了该服务的客户被错误地预测为没有订购该服务，值 165 表示模型正确预测的 165 个订购了该服务的案例。

从混淆矩阵包含的四部分信息能够计算出由混淆矩阵衍生出的准确率、敏感度、特异度、正例命中率、负例命中率，数值如下：

准确率：(TP + TN)/总数 = (165 + 6548)/(165 + 6548 + 214 + 573) = 0.8950

召回率：TP/(TP + FN) = 165/(165 + 573) = 0.2266

误报率：FP/(FP + TN) = 214/(214 + 6548) = 0.0316

精确度：TP/(TP + FP) = 165/(165 + 214) = 0.4353

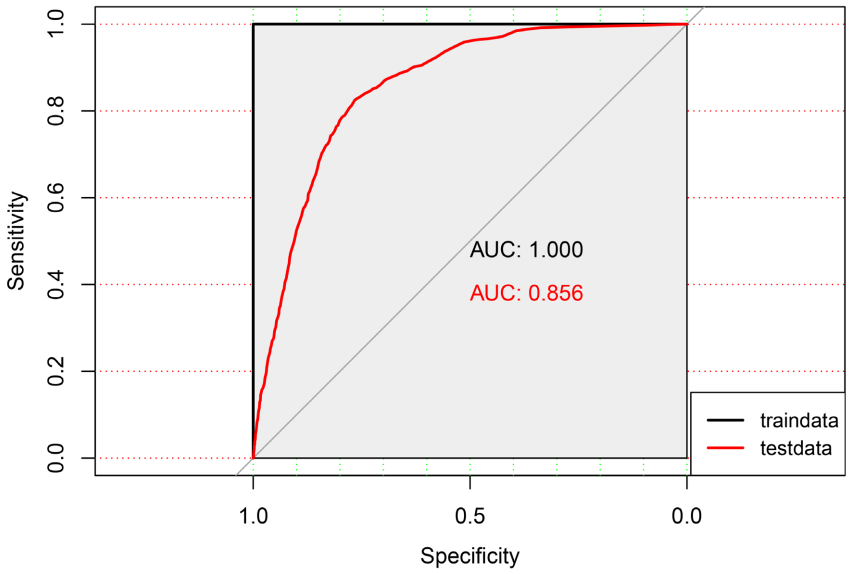

Figure 6. ROC curve of random forest
图 6. 随机森林 ROC 曲线

图 6 表示的是随机森林模型根据训练集和测试集生成的对应 ROC 曲线[14]。从图中可以看到训练集的 AUC 显示为 1.000 且 ROC 曲线完美贴合区域边界，这说明对于训练集数据而言，模型能够捕捉所有能区分样本是否订购该服务的信号，表示在训练集上模型拥有完美的预测效果。测试集的 AUC 显示为 0.856 并且 ROC 曲线较于训练集没有非常贴合区域边界，说明模型对于测试集数据而言，能够捕捉大部分区分样本是否订购该服务的信号，但还有少部分的信号未能捕捉到。从测试集的 ROC 曲线可以看出，相较于训练集而言模型的性能有所下降，这种差异可能是由于模型在进行训练时学习了训练集中的特定特征，但是这些特征在测试集中没有表现出来。

随机森林模型基于测试集数据生成的 ROC 曲线揭示了模型拥有良好的分类预测能力，能够较好地对数据进行分类。虽然随机森林模型预测准确率高达 89.5%，但是它在精确度和召回率上的表现还有所欠缺。低水平的召回率可能会导致在实际应用中给银行机构带来较大的损失，因为会使得银行错失大量最终会订购该产品的客户。

3. 模型比较分析

根据表 7 与图 7 可知，在三种模型当中，随机森林的 AUC 值最高说明其拥有最好的分类性能，Logistic 回归模型的 AUC 值第二高说明其分类性能次于随机森林，支持向量机的 AUC 值最低说明其分类性能最

差。但这三种模型的预测准确率都在 90%附近，结合三种模型的 AUC 值推测出现这种情况的原因是由于测试集中样本分类不均衡，未订购的样本数远大于订购的样本数。

Table 7. Evaluation indicators of model classification performance
表 7. 模型分类性能评价指标

	准确率	精确度	召回率	误报率	AUC
logistic 回归	90.50%	7.33%	45.10%	91.20%	0.794
SVM 模型	90.79%	66.25%	7.30%	99.60%	0.536
随机森林	89.50%	43.53%	22.60%	96.80%	0.856

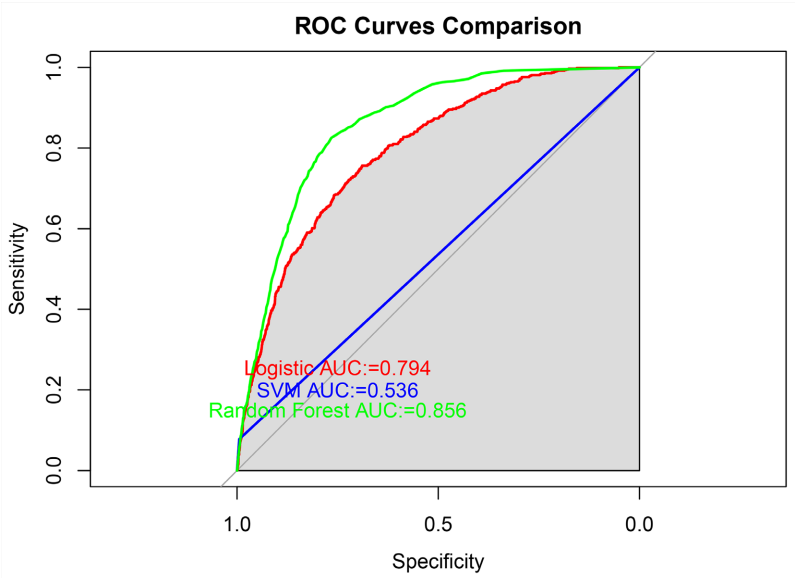

Figure 7. ROC curves of each model
图 7. 各模型 ROC 曲线

5. 总结与展望

1. 模型性能比较：通过对随机森林、Logistic 回归和支持向量机三种模型的比较，发现随机森林具有最佳的分类性能，其次是 Logistic 回归，而支持向量机表现最差。这表明随机森林在银行客户订购金融产品的分类问题上具有较强的预测能力。

2. 客户特征影响：从 Logistic 回归模型中得出了许多关于客户个人特征和行为的结论。例如，年龄、职业、婚姻状况、教育水平等因素对客户是否订购金融产品有显著影响。具体而言，年龄越大、通话时间越长、活动联系次数越多等因素都会增加客户订购金融产品的可能性。

3. 经济因素影响：经济指标也在一定程度上影响着客户的金融产品订购行为。例如，就业变动率、消费者价格指数、消费者信心指数等指标都与客户订购行为密切相关。这表明经济环境的不同方面对客户的金融产品选择产生了影响。

4. 建议和展望：最后，报告提出了一些针对银行营销策略和产品推广的建议。例如，建议银行制定针对不同年龄段、职业群体的个性化营销策略，以及在经济环境变化下及时调整营销策略和产品设计。

综合以上结论，可以得出在制定银行营销策略和产品推广方面，需要考虑客户个人特征、行为指标和经济环境等多方面因素的影响，并制定相应的个性化和针对性策略，以提高金融产品的销售效果。

参考文献

[1] 陈良凯, 黄登仕. 消费者银行理财产品购买行为实证研究[J]. 西南金融, 2013(10): 33-36.

[2] 屈忠锋, 吴鸿华, 李凡军. 基于 Logistic 回归的中小企业信贷风险评估与信贷策略优化建模[J]. 山东大学学报(理学版), 2024, 59(8): 94-102.

[3] 谭小燕, 柳燕, 孙情情, 等. 基于Logistic回归分析的行为干预对肾结石病人术后复发的影响[J]. 循证护理, 2024, 10(9): 1642-1646.

[4] 祝珊珊, 马晶林. 基于支持向量机的学分银行构建探究[J]. 新疆开放大学学报, 2023, 27(4): 65-71.

[5] 张微, 刘云杰, 张舒婷. 基于随机森林与神经网络算法的个人信贷智能决策系统研究以德国某银行的共享样本集为例[J]. 经营管理者, 2024(3): 94-95.

[6] 中国人民银行固原市分行课题组, 何文虎. 基于 Logistic 回归模型的涉农贷款风险影响因素研究——以宁夏固原市某农村商业银行 5584 户涉农贷款户为例[J]. 黑龙江金融, 2024(2): 84-88.

[7] 吴尚智, 王旭文, 王志宁, 等. 利用粗糙集和支持向量机的银行借贷风险预测模型[J]. 成都理工大学学报(自然科学版), 2022, 49(2): 249-256.

[8] 金朝亮. 基于机器学习的银行商户信用评估研究[D]: [硕士学位论文]. 银川: 宁夏大学, 2023.

[9] 楚东方. 基于支持向量机的小微企业贷款违约风险预测研究[D]: [硕士学位论文]. 成都: 西南财经大学, 2022.

[10] 孙凯旗. 基于 SVM 模型的上市商业银行系统性风险预警研究[D]: [硕士学位论文]. 重庆: 重庆工商大学, 2022.

[11] 刘恬恬. 企业资产支持证券价格影响因素的实证研究[D]: [硕士学位论文]. 天津: 南开大学, 2022.

[12] 刘香. 基于 SVM 模型的农村金融机构农户信用风险评价体系研究——以黑龙江省为例[J]. 市场周刊, 2024, 37(7): 19-24.

[13] 曹桂林, 杨许亮, 王若凡. 基于机器学习的银行客户流失分析[J]. 山东商业职业技术学院学报, 2024, 24(1): 105-110.

[14] 陈衍姣. 基于改进的随机森林算法的绿色信贷信用风险评估研究[D]: [硕士学位论文]. 贵阳: 贵州财经大学, 2023.