
商业银行营销策略分析

杨谨行、段颖凡、宋俊儒¹

摘 要

电话营销是一种低成本、高效率的营销方式，自 1970 年问世以来备受大型企业的青睐。

本案例分析了商业银行对客户进行电话营销以吸收存款的数据，首先采用 Logit 回归模型、支持向量机和决策树三种方法对营销数据建模，找到影响营销成功的因素，并预测潜在客户，帮助商业银行更合理地分配资源、节约成本，同时提高利润。其次，对营销成功客户进行分类和画像，以针对不同人群，找出其特点和最适合的营销策略。

与其他使用同一或类似数据集的研究不同在于，我们在前期使用了大量的可视化结果来支持特征选择和数据处理、对每个变量与因变量和各个变量之间互相交叉，挖掘了指标间的潜在联系。其次，我们采用了不同的模型：一个高召回率的决策树模型，适合银行采用“广撒网多捞鱼”的策略，确保潜在客户的营销范围覆盖；一个高精确率的 SVM 模型，适合银行经费在有限情况下追求每通电话都有效挖掘成功客户；表现均衡的 Logit 模型，适合给出稳定的预测，方便企业根据其成本收益的实际情况来选择营销手段。在通过历次营销数据提供预测模型基础上，我们还进一步运用聚类算法对潜在客户群体进行了画像分析，进一步帮助金融企业进行客户关系管理和存贷款业务拓展。

最终，本文得到的结论是，针对性地对客户进行营销活动、根据不同客户群体推荐适当的产品、在营销人员培训上保证其营销方法和技巧，则能在控制成本的情况下，提高营销效率，为银行吸纳更多的客户和存款。

关键词：电话营销、Logit 回归模型、支持向量机、决策树、不平衡样本采样处理、两步层次聚类、银行客户用户画像、随机森林缺失插补

¹ 联系方式：

杨谨行 2017312292 应用统计学 17 yang_jinhang@163.com 13120496200

段颖凡 2017310911 应用统计学 17 yingfan_duan@163.com 18690413666

宋俊儒 2018310881 统计学 18 2018310881@email.cufe.edu.cn 13080751013

目 录

1. 前文介绍	3
1.1 案例背景	3
1.2 研究对象	3
1.3 文献综述	3
2. 数据准备方案	4
2.1 数据来源	4
2.2 数据说明	4
2.3 数据基础清洗	5
2.4 探索性数据分析与特征处理	6
2.4.1 个人信息类特征	6
2.4.2 本轮营销情况	10
2.4.3 上轮营销情况	12
2.4.4 经济及宏观环境变量	13
2.5 数据处理	14
2.5.1 统一变量类型	14
2.5.2 数据缺失处理	14
3. 模型及方法	16
3.1 预测模型	16
3.1.1 Logit 回归模型	16
3.1.2 支持向量机模型	18
3.1.3 决策树模型	19
3.1.4 模型比较	22
3.2 聚类模型	23
3.2.1 聚类及用户画像综述	23
3.2.2 营销成功客户群体的聚类	25
3.2.3 营销失败客户群体的聚类	25
4. 结论	27
参考文献	28

1. 前文介绍

1.1 案例背景

商业银行是管理货币基金的特殊企业，需要在维持运营的时候最大化扩充客户群体。传统的商业银行的来源主要是以较低的利率借入存款、以较高的利率放出贷款获得的利差。

相关的市场营销理论（Marketing Theory）始于一般工商企业市场，由威斯康星（Wisconsin）首次提出；后来哈佛（Harvard）学派主要贡献在于对市场营销问题的编辑和整理；后来，市场营销系统过程理论和消费者中心理论相继问世[1]。商业银行的市场营销理论研究则是在 1958 年全美零售银行联合会会议上第一次提出银行市场营销概念，标志着西方银行市场营销理论正式诞生，众多手段中，电话营销是众多营销手段中，低成本、高效率获客的代表[2]。最早的电话营销出现于 20 世纪 70 年代的美国，由免费电话的产生开始，随后逐渐发展到亚洲，并在 90 年代在中国大陆得以发展。随着电话等通信手段的普及和以客户为中心的买方市场的形成，电话营销在很长的一段时间里被越来越多的现代企业所采用[3]。

1.2 文献综述

前人在商业银行电话营销方面已经有应用数据挖掘技术用来预测经过电话营销销售银行长期存款的结果。1956 年，美国著名的市场学家 Wendel R. Smith 首次提出市场细分的概念。60 多年来，市场细分的理论和方法得到不断完善，并被广泛应用于营销实践。对于商业银行而言，其拥有广大客户群体，且客户之间存在较大差异，市场细分对其提高服务水平具有重要意义。在数据挖掘中，聚类分析是用来实现客户细分的主要方法之一[28]。H.W.Shin 等于 2004 年采用 K-means、SOM（self-organizing map）、fuzzy K-means 聚类算法对股票交易客户进行细分[29]；Hany 在 2014 年应用逻辑回归、基于 C5.0 的决策树、贝叶斯网络、多层感知器神经网络在银行营销数据集上，并利用多种指标比较这几种模型的效果，他发现 C5.0 算法下的决策树略优于其他模型，并且营销时长对于成功率有很强的影响[4]。针对同一数据集，Safia Abbas 在 2015 年也使用了决策树(DT)和粗糙集(RST)理论进行了建模预测和分析，高丽在 2017 年应用逻辑回归、决策树和支持向量机用来预测电话营销的效果，可以识别出较高价值的客户，避免在低价值客户身上浪费资源，以此提高投入产出比[5]。

2. 数据准备方案

2.1 数据来源

本案例数据源来自 UCI 数据库，原始数据采集于葡萄牙某银行，该银行在针对存款业务推广的市场营销活动中收集了部分客户数据，目前，该数据集有新旧两个版本，每个版本有抽取后数据和原始数据，我们选择了最新最全的数据集进行分析²。

2.2 数据说明

数据集中共有 79 个变量，下表中对本节及第三小节模型中涉及的 18 个主要变量进行了说明：

表 2-1 数据变量说明表

变量类别	变量名	变量符号	类型	变量含义	变量取值范围
个人信息	年龄	age	连续	营销对象的年龄	[17, 98]
	职业	job	离散	营销对象的职业	housemaid、services、admin 等 12 类取值
	婚姻状况	marital	离散	营销对象的婚姻状况	single、divorced、unknown 等 4 类取值
	教育情况	education	离散	营销对象的教育情况	basic.4y、high.school 等 8 类取值
	违约	default	离散	营销对象的违约状况	no、unknown、yes 等 3 类取值
	房屋贷款	housing	离散	营销对象的房屋贷款	no、yes、unknown 等 3 类取值
	其他贷款	loan	离散	营销对象的其他贷款	no、yes、unknown 等 3 类取值
	联络方式	contact	离散	本次营销的联络方式	telephone、cellular 等 2 类取值
本次营销情况	联络月份	month	离散	本次营销的联络月份	may、jun、jul、aug 等 10 类取值
	联络日	day_of_week	离散	本次营销的联络日	mon、tue、wed 等 5 类取值
	联络时长	duration	连续	本次营销的联络时长	[0, 4918]
	联络次数	campaign	连续	本次营销的联络次数	[1, 56]
上次营销情况	上次联络间隔	pdays	连续	距上一轮营销的间隔天	[0, 999]
	上次联络次数	previous	连续	之前总共的营销次数	[0, 7]
	上次营销结果	poutcome	离散	上次营销是否成功	nonexistent、success 等 3 类取值
经济及社会环境	员工变动率	emp.var.rate	连续	季度就业变动变动率	[-3.4, 1.4]
	消费价格指数	cons.price.idx	连续	月度消费者价格指数	[92.201, 94.767]
	消费者信心指数	cons.conf.idx	连续	月度消费者信心指数	[-50.8, -26.9]
	伦敦三月拆借利率	euribor3m	连续	伦敦隔夜拆借利率	[0.634, 5.045]
因变量	员工总数	nr.employed	连续	公司季度员工数	[4963.6, 5228.1]
	本次营销结果	y	离散	客户是否下定金	no、yes 等 2 类取值

数据中的 11 列离散型变量分布如下图所示：

² UCL 对数据集的命名为 bank-additional full.csv

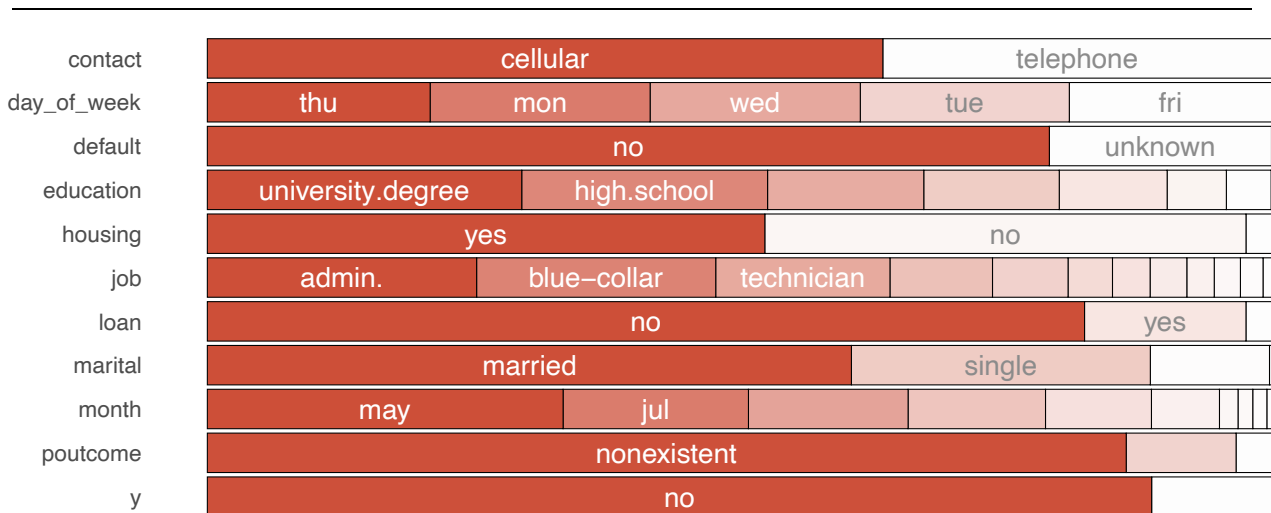


图 2-1 离散型变量分布图

可以看到，存在一些 nonexistent, unknown 等标记缺失的变量，需要后续处理。

数据中的 10 列连续型变量分布如下图所示：

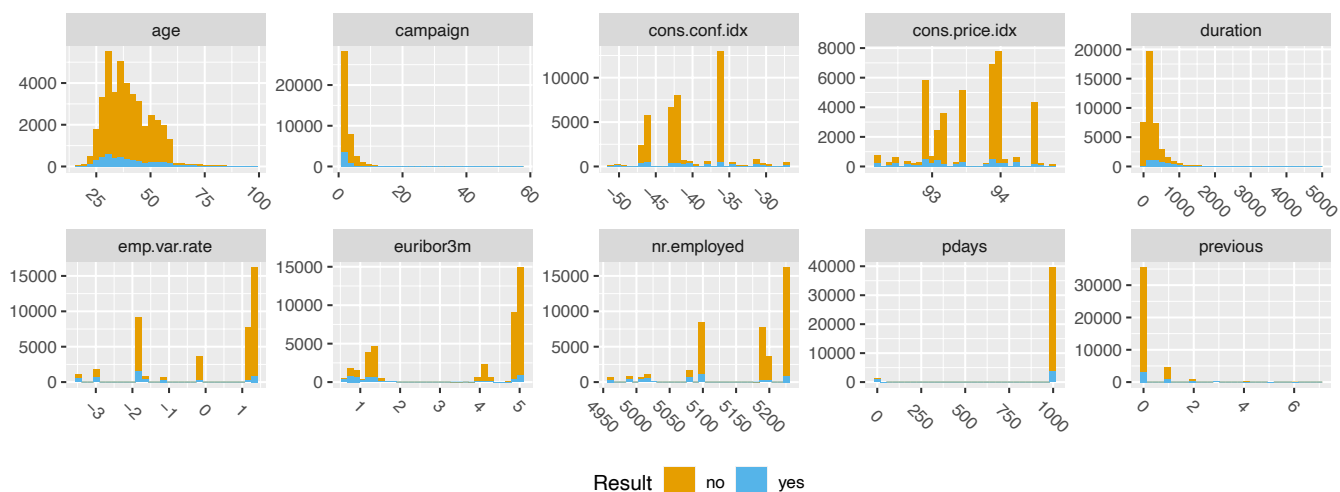


图 2-2 原始数据连续型变量分布图

从上图可以看出，不少变量的取值不是完全连续的，经济变量都只有个别值。

简单浏览数据情况后，我们接下来将具体分类讨论，每个数据的具体处理方式。

2.3 数据基础清洗

依照 2.2 节的发现，我们先对数据进行了一些简单的清洗和转义：

- **数据冗余：**在数据的 41188 行中，有 12 行数据出现了完全一样的数据。本数据没有客户 ID 等主键信息，所以无法使用主键重复等方式判断数据冗余，但考虑到完全一致概率较小，有可能是数据重复录入导致，因此进行了去重处理；
- **数据转义：**首先，数据集虽然没有缺失，但存在不少对 NA 的标记方式，如在 pdays 变

量中 999 标记缺失，在 job、marital、education、default、housing、loan 中用 unknown 代表缺失，为方便后续分析和缺失填补，预先处理为 NA 值；其次，月份、工作日等也用了缩写编码，也处理成相应的数字。

2.4 探索性数据分析与特征处理

在完成了初步的变量理解和基础清洗后，2.4 部分将用较大篇幅具体介绍我们对每个变量做的特殊清洗工作，并用可视化的方式加以说明处理的原因。

2.4.1 个人信息类特征

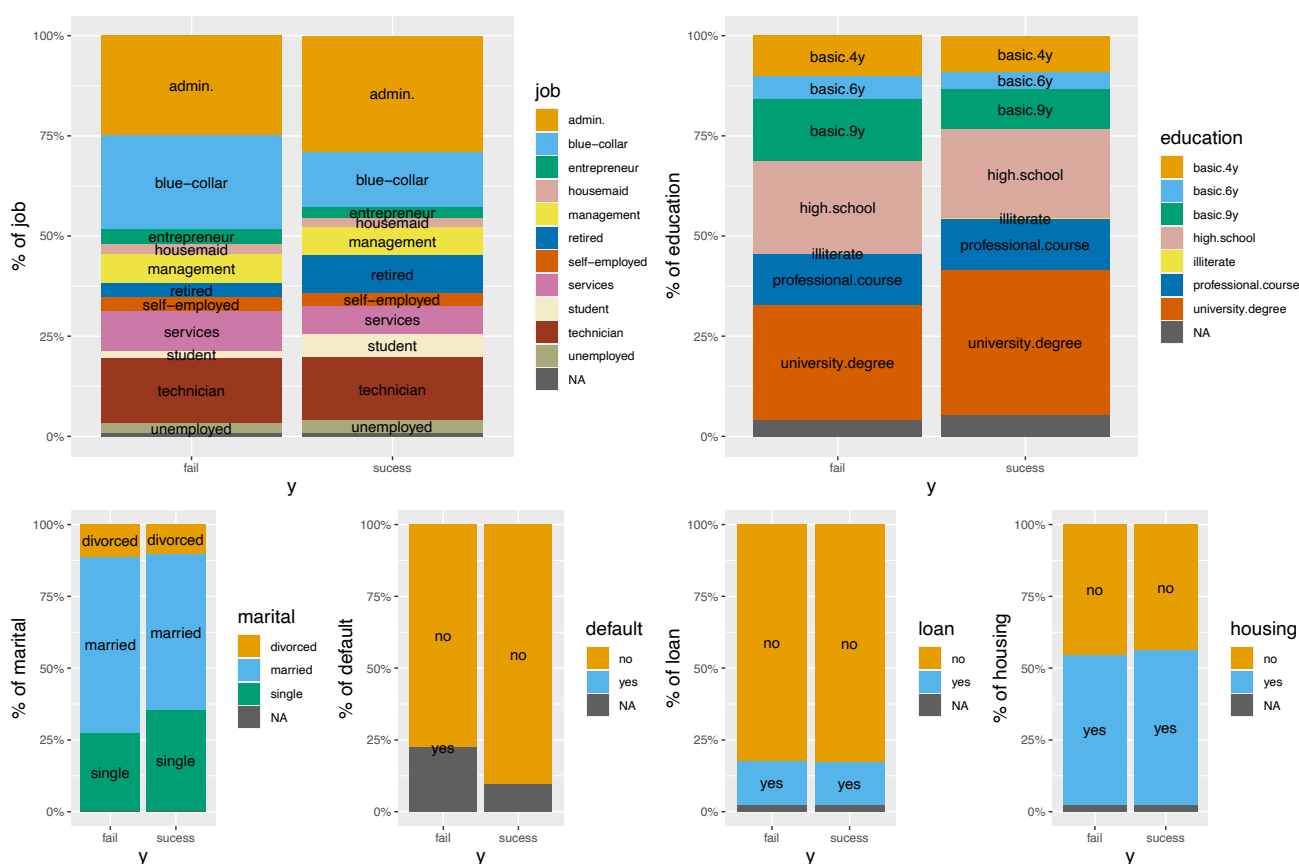


图 2-3 个人信息类属性变量一览图

通过上图可以初步看出，单身对营销成功有一定影响，而贷款影响程度不大；年龄、职位和教育年限中，蓝领、服务业人员更多选择拒绝营销，而学生和白领更多选择接受，然而，这只是变量与因变量的二维度关系，下面我们进一步考虑多个变量的相互作用，得出了以下模型指标的选择和处理方案结论：

1. Default 变量的缺失视作一类，并重新分为没有违约记录和其他

在图 2-3 中我们已经看出 default 变量的缺失，会显著的影响营销的结果，推测很可能是

因为缺失信用数据的可能没有银行卡的使用习惯，从而导致营销的成功概率降低。此外，数据中只标记了 3 个有违约记录的用户，为避免该特征在 one-hot 之后该列取值大部分为 0，将该列重新划分为没有违约记录和其他，执行操作的前后的 mosaic 图变化如下所示：

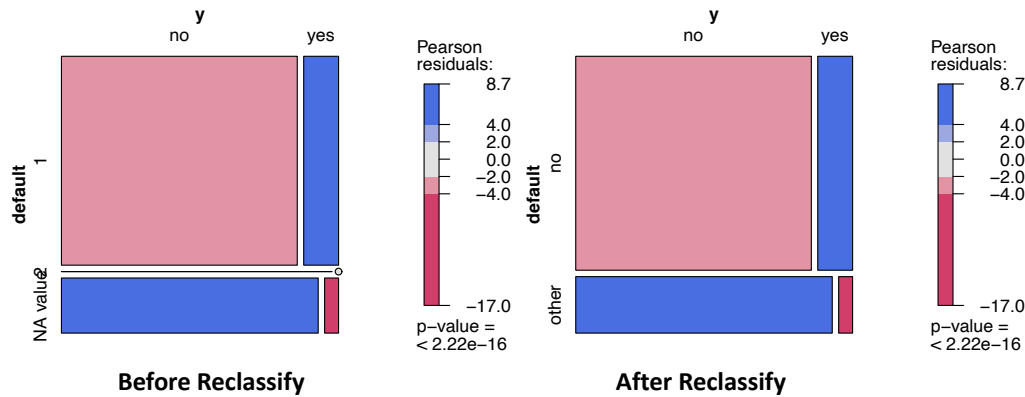


图 2-4 default 变量处理前后对比图

通过上图可以看出，明确无违约记录的客户相比其他客户，接受营销的概率显著增加。

2. 教育类型离散型变量转为教育年限次序型变量

由于教育的年份在变量标签中已经给出三个，结合葡萄牙高中、大学的年限信息，我们可以将被营销对象的受教育程度用教育年限而不是教育类型来表示，特征转化后前后，分别与职业变量交叉，结果如下图所示：

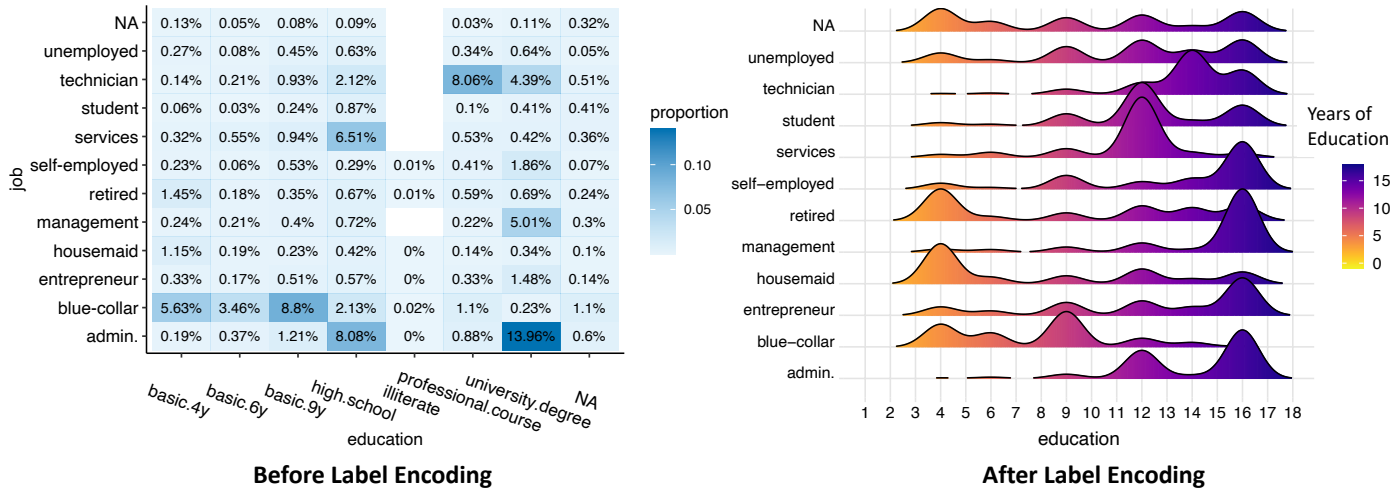


图 2-5 education 变量处理前后对比图

从左图中可以看到，拥有大学学历的行政人员在被营销人员中占比最大，另外、蓝领工人的学历集中在基础 9 年教育及以下，技术人员多为职业教育课程或大学学历，都符合我们的常识认知，因此对没有教育、高中、职业课程、大学学历分别赋值为 1、12、14、16，可以将特征转为右图的峰峦图（分组密度曲线）所示，从而避免让 education 变量产生过多哑

变量影响模型稳定性。

3. 结婚与离婚的对因变量几乎无影响，应当归为一类

尽管结婚与离婚的营销对象可能在生活习惯上有很大差异，但数据表明这两类没有分开的必要。

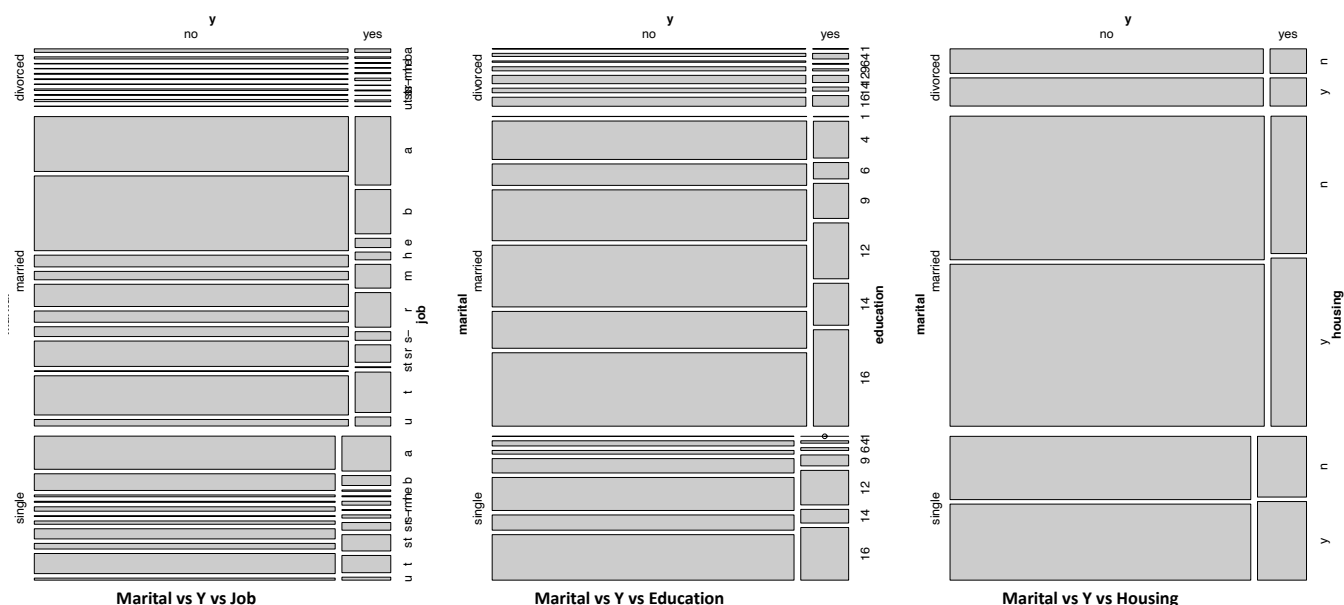


图 2-6 Marital 变量处理前示意图

上图中，我们将 Marital 随机选择了 Job、Education、Housing 三个变量（其他变量结论一致）分析 Marital 的三类，在进一步考虑不同职位、教育、贷款情况下，是否对 y 的影响有差异。而上图的三张子图中，先看左侧分类标准、再看上侧分类标准，我们可以看到 married、divorced 和 single 的三类下，single 会显著造成影响结果的成功概率增加，但 married 和 divorced 在这三种效应的交叉下比例都非常接近，尽管我们在看前两张图的左侧和右侧分类标准的时候可以清晰地看出他们的职位、教育都有显著的不同。

基于这种考虑，我们将 Marital 重新划分为两种类型，结过婚（用 1 表示）和没结过婚（用 0 表示），处理后结果如下图所示：

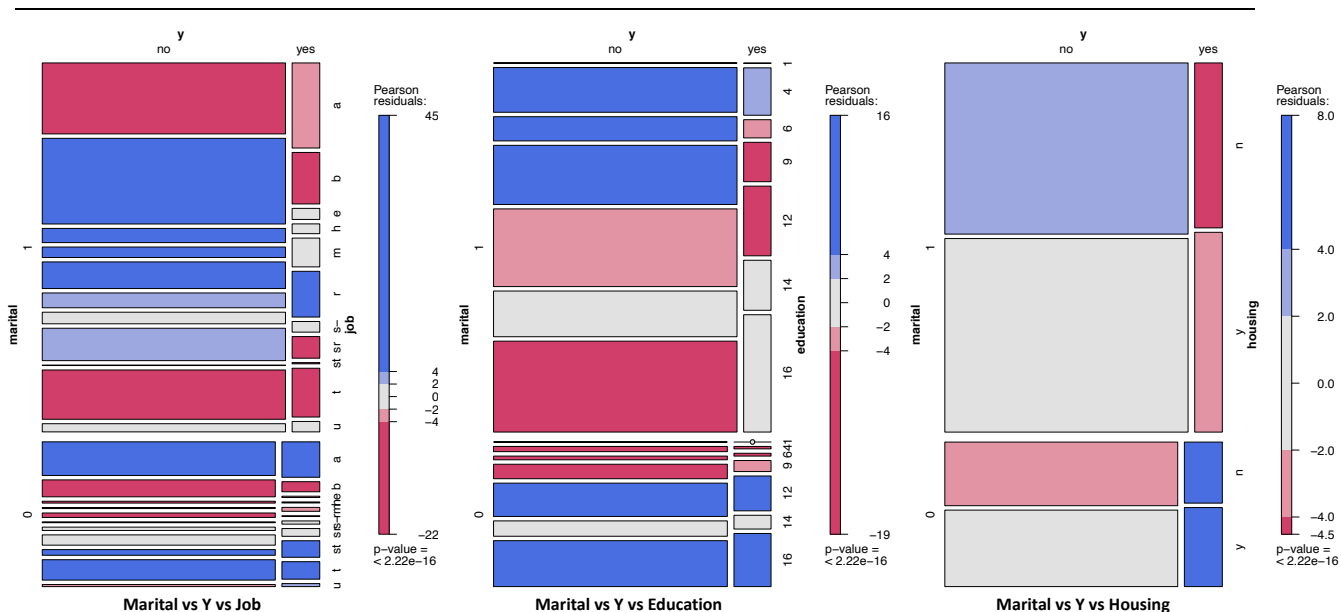


图 2-7 Marital 变量处理后示意图

可以看到，在 Job、Education 和 Housing 三组交叉维度下，Marital 变量的 0、1 可以显著影响营销的成功与否，且该变量不需要进一步 one-hot 生成更多的 01 变量。

4. 保留年龄变量的数值型类型不做分箱

对年龄进行因子分箱，用年龄段的是一种常见的提升模型稳定性和解释性的常见做法，但在本数据集中，我们认为这样的操作会带来过大的信息损失：

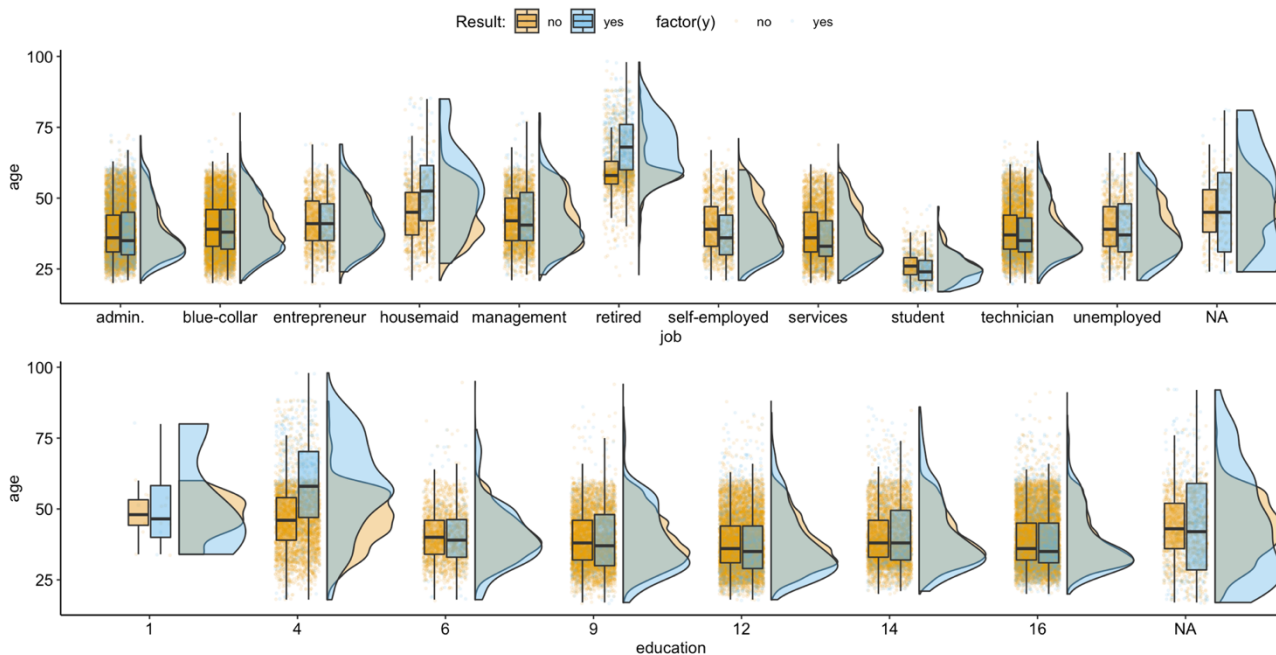


图 2-8 age 按职业和教育分组云雨图³

³ 该图的每一个小图由三部分构成——分组箱线图、分组散点图、分组半小提琴图，可用于分析数据的分布、离群、均值及分位数等特征

从上面的云雨图中可以看到，不同的年龄层之间，并不能确定一个划分的依据，如上侧子图中，数据中尽管很多职位的年龄峰值都集中在 30 岁左右，但是退休人员的年龄分布在 65 岁之上，且年龄越大反而越容易接受营销，其他的职业都是营销成功的对象平均年龄会更低。下侧子图中，教育年限为 4 年的人员也有类似的偏差。

因此，年龄应该保持为连续的数据类型。

与此同时，而两个子图的缺失值年龄特征相近，也提示我们后续可以利用 `age` 和其他信息来对其缺失进行预测填补。

2.4.2 本轮营销情况

1. 基于 `month` 提取出年份和月通话量特征

观察原始数据集，我们发现观测值是通过时间顺序记录的，因此结合月份标签，我们首先提取了年份特征，将其月份一起考虑，我们将观测分层如下：

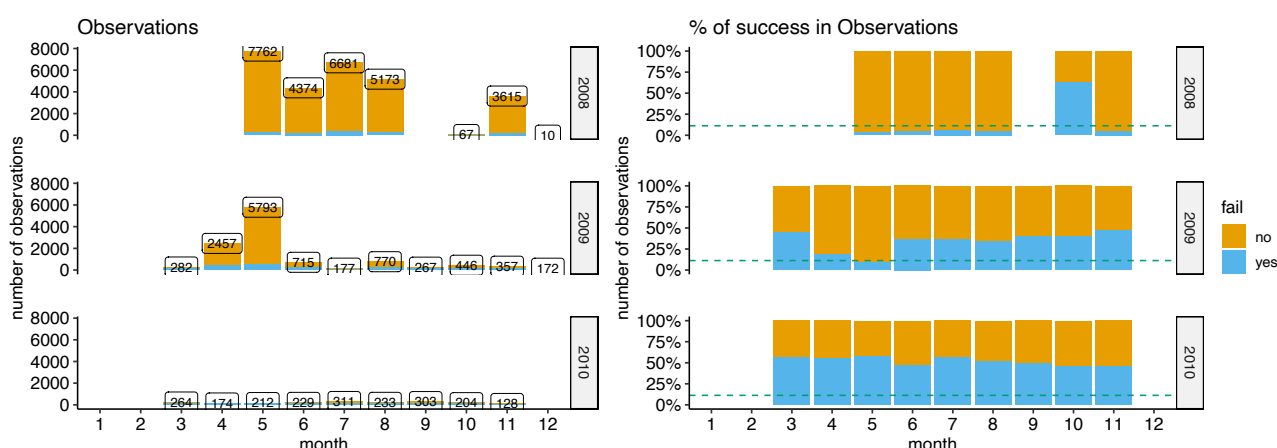


图 2-9 观测按年份与月份划分直方图

可以看到样本在每年前 2 个年都不进行营销，而 2018 年的 9 月也没有营销的数据。

同时，有一个极为特殊的现象出现：似乎成功率的高低，与每月电话打出的数量高度相关。最典型的是 2009 年的 3、4、5 月，随着拨打电话数量的增加，成功率逐渐递减，基于这个发现，我们进一步提取每月通话量特征：

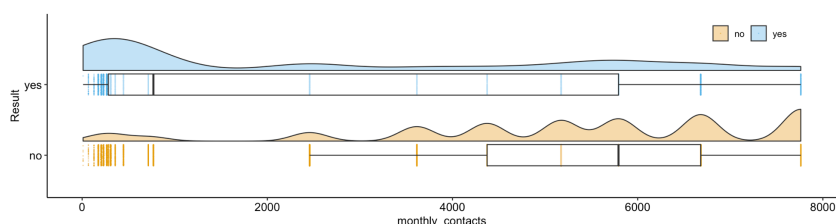


图 2-10 每月电话数量云雨图

从这个变量我们可以显著看出，26 个不同的年、月数据显示成功的月份集中在拨打电话为少数的几个月，而没有成功营销的数据大多集中在拨打电话密集的月份，这非常有可能是营销人员因为工作任务加大，工作质量降低导致的。

2. 删除不显著的上次通话星期变量

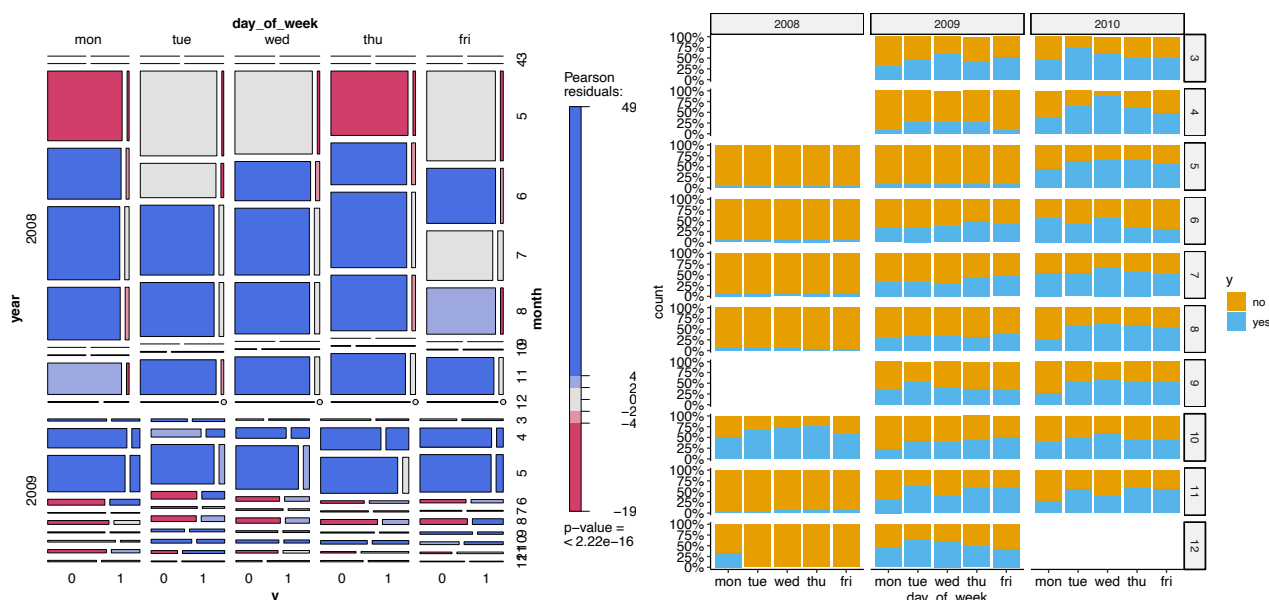


图 2-11 星期变量分布图

从上图也可以明显看出，在不同的月份、年份下，并没有固定的规律，比如 2019 年的周三在三月份的时候最高，但是在 10 月份又是最低。这是因为，不同的月可能有节日因素的影响，而且这个变量本身并不是播出这次电话的星期，而是上一次的，因此没有太大的预测价值，不将此变量引入模型。

3. 保留 contact 变量但删除 duration 变量

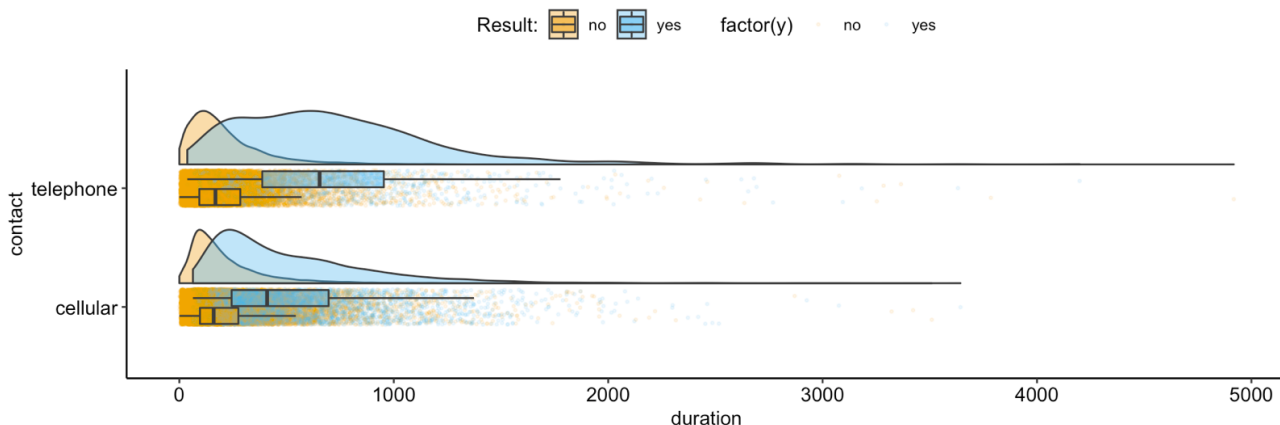


图 2-12 duration 按电话类型分组云雨图

从图中可以看出，电话时间的延长对获客的概率增加显著提高，但这个变量存在三个问题：因果关系不明确（可能是营销成功的客户进一步通话继续了解）、变量获取顺序的问题（打电话前，并不知道本次通话的时间，因此不能做预测变量）、分布不均衡（有通话时间在一个小时以上的变量）；但另一方面不同的联络方式下使用手机的通话时间整体更长，且营销成功概率更大。另外该变量可以在拨打电话前确认，因此予以保留。

2.4.3 上轮营销情况

1. 删除含义不明确的 pdays 变量

在数据集说明中，pdays 的 999 表示距离上轮营销的天数缺失、campaign 取 1 表示客户是在这一轮中第一次联系，但 pdays 的缺失值为 39661，而 campaign 取 1 的仅为 17634，基于数据质量和模型解释性的考虑删除 pdays 变量。

2. previous 与 poutcome 只保留其一

Previous 取 0 和 poutcome 取 nonexistent 都表示是第一次联系，经过检查，两者的 35551 行观测是一一对应关系。

而在其他的 5625 个观测下，我们看到，不管过去营销成功了多少次，只要上次成功，则本次成功的概率大大增加，而不同的拨打电话次数下，平均的成功概率还会来回波动，这可能是因为可能上次营销拨打电话次数太多，客户感到厌烦，但同时多次拨打也能保证跟客户联系紧密，因而导致 previous 变量的 1~7 取值对 y 并不是一个线性的贡献，应该当作分练变量处理，然而 poutcome 变量已经囊括了绝大部分数据的情况（previous=0），因此综合考虑将 previous 变量删除，只保留 poutcome 的 0 和 1 作为客户历史营销记录的标记。

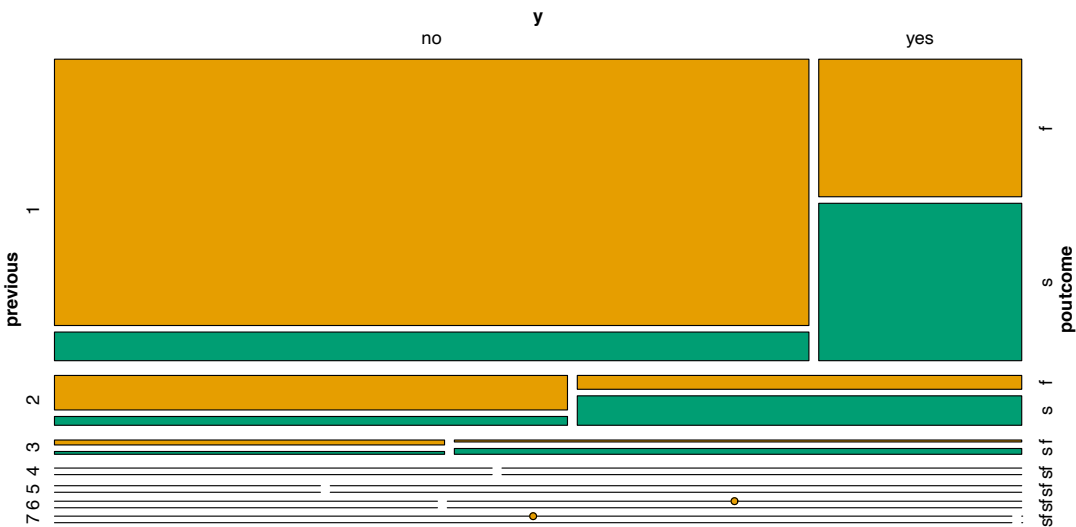


图 2-13 曾经营销成功客户次数与结果对比图

2.4.4 经济及宏观环境变量

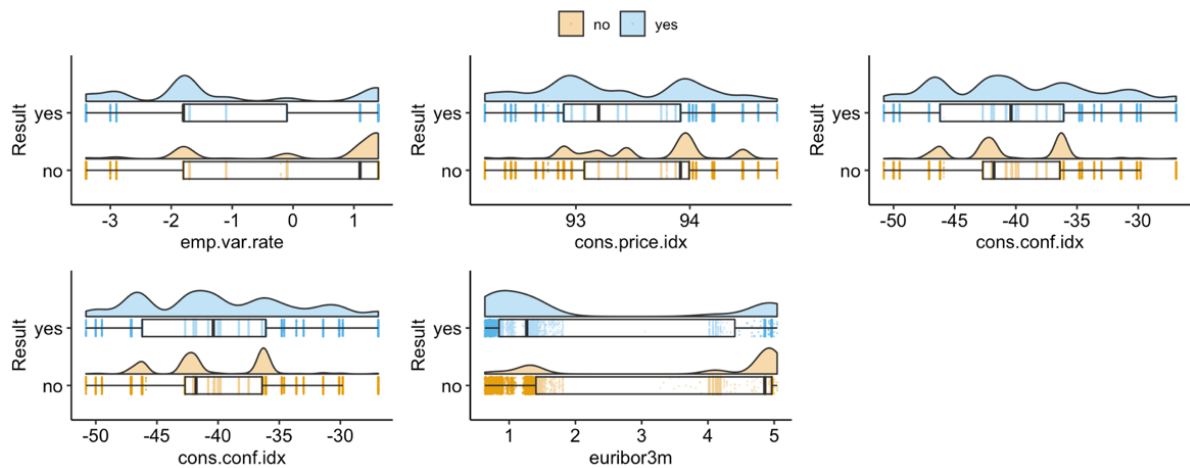


图 2-14 经济指标与营销成果关系图

数据集中的宏观经济变量以面板数据的形式给出，如上图所示，不同的经济背景对营销有一定的影响，如隔夜拆借利率在 5 的时候，经济环境整体资本机会成本增高，客户可能更倾向于减少消费增加存款，因而营销效果不理想。

然而，结合我们之前提取的年份数据，我们可以进一步对数据进行时间序列的可视化，发掘特征信息：

1. 对指标进行标准化处理去除量纲

下图左图中可以看到，所有的指标的波动程度差不多，但量纲不同，导致纵坐标有很大差异，因此需要做均值归一化处理。从结果中可以看出 CPI 和员工变动率的波动情况基本保持一致。此外，考虑到时间序列的预测上，没有给个体标签，而且相关指标主要在经济层面上：

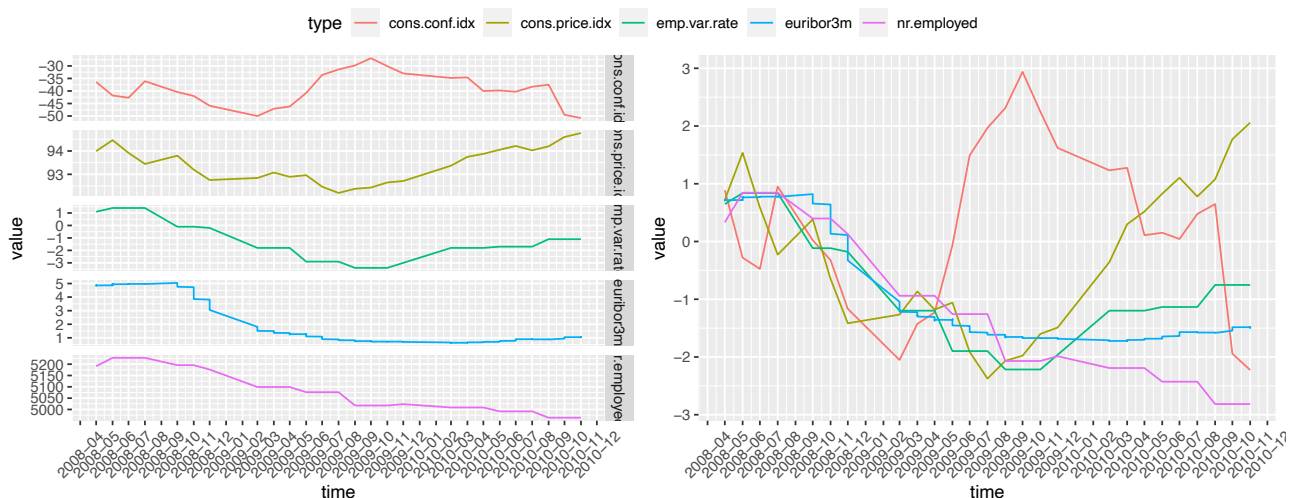


图 2-15 经济变量均值处理前后对比图

均值归一化处理完之后的结果如右图所示，可以看到员工人数每年都在递减，而消费者信心指数在 CPI 下跌的时候会与其反向提升。

2.5 数据处理

依照 2.4 节提取特征信息、对特征层面进行了处理后，我们对数据的观测进一步进行处理。

2.5.1 统一变量类型

我们对数据的类型进行了统一，转化为 numeric 和 factor 类型，依照其分组成功率以此排 factor 的顺序，以便后续做哑变量标记。

2.5.2 数据缺失处理

经过对变量的清洗后数据集的缺失情况如下

表 2-2 数据变量说明表

每个观测缺失的变量数	无缺失	1	2	3	4
行数	30378	9032	1282	231	41

每个变量缺失观测数	Job	Martal	Education	Housing	Loan	其他变量
行数	199	71	1595	988	988	无缺失或缺失已通过特征变换处理

1. 缺失值填补

在缺失值填补上，我们一共执行了三个基础方案，并对其进行混合。

表 2-3 缺失值填补方案

方法	Logit 模型平均 accuracy	SVM 模型平均 accuracy	决策树模型平 均 accuracy
众数/均值填补	28.2%	50.6%	32.2%
Pmm 均值预测插补	34.5%	78.5%	36.5%
随机森林	38.4%	74.7%	35.4%
综合方法	39.5%	77.4%	38.2%

其中我们设计的综合方法具体的方式为：

- 针对离散型变量：随机森林填补结果和均值填补结果冲突较大时，参考众数；冲突较小时，针对不同模型使用最适合它的结果；
- 针对连续型变量：使用不同插补方法的准确率倒数作为权重，加权平均出一个综合的填补值（后续因 pdays 等变量的删除，连续型变量没有继续出现缺失）。

2.少量删除缺失观测

在缺失值中，housing 和 loan 的缺失保持一致，education 缺失较多，为主要预测项，但它的预测很大程度依赖同类变量的 Job 和 Marital，因此我们对 Job、Marital、Education 三个指标全部缺失的 27 行观测进行了删除操作，以保证其因为缺失填补的不稳定性导致了结果不统一。

3. 模型及方法

3.1 预测模型

本节通过建立模型对因变量进行预测，采用的模型有：Logit 模型、支持向量机模型和决策树模型，三者进行效果比较，对营销策略进行分析。建模过程中使用 70%数据作为训练集，30%数据作为测试集。

3.1.1 Logit 回归模型

1. 模型介绍

Logit 模型是广义线性模型的一种特殊形式，它通过连接函数，对因变量 y 属于某一类的概率建模而非直接对 y 建模[7]。本文采用的 Logit 回归模型为： $\ln(\frac{\pi}{1-\pi}) = \alpha + \beta X$ ，其中， $\pi = P(Y = 1|X)$ ，表示营销是否成功。Logit 模型再分类问题上有非常好的应用，因为它不仅可以因为它不仅可以给出每个自变量对因变量有什么样的影响，还可以对新进入的样本进行较为准确的预测[8]。

2. 模型建立

根据描述性分析，数据集的因变量 0/1 分布非常不均衡，而逻辑回归采用的是经验风险最小化作为模型的学习准则，因此它的目标是使得训练集总体的后验概率最大。因此这种算法会将关注点更多地放在多数类的拟合情况下，多数类别的分类正确与否，更为影响最终整体的损失情况。而在本文的预测任务中更关注的是少数类别的分类正确情况，即营销成功的正确情况。对训练集的 0/1 分布进行检验，发现 94.5%的数据的因变量为 0，因此对训练集的数据进行重新采样。本文一共采用了四种方法：过采样，欠采样，两者的结合和 SMOTE 法，然后分别在测试集上比较效果，测试集上四种方法的预测结果如下表所示。

表 3-4 抽样方法效果比较表

	accuracy	precision	recall	F1
Over sampling	0.6812312	0.8372911	0.7170217	0.7725033
Under sampling	0.6873223	0.8405054	0.7229395	0.7773022
Both	0.6784699	0.842031	0.7065849	0.7683847
SMOTE	0.257289	0.9408284	0.01710781	0.03360457

考虑到测试集中的因变量同样分布不均，采用 $F_1 - measure$ 值作为评价指标，可以看到

欠采样的 $F_1 - measure$ 最高，其次是过采样，两者相差不多，但由于欠采样对于样本数据浪费过于严重，我们选择过采样后的训练集进行后续模型建立。

本节首先在描述性分析的基础上建立 Logit 回归模型得到全模型，并基于 AIC、BIC 准则对模型进行简化，分别得到 AIC 模型和 BIC 模型，系数估计和标准误差如表所示，从表中可以看出，就 AUC 而言，BIC 模型的效果仅略微弱于 AIC 模型和全模型，并且 AUC 最大，因此我们选择 BIC 准则下更为简洁的模型作为最终模型。

表 3-2 Logit 模型回归结果

变量	全模型		AIC 模型		BIC 模型	
	系数	标准差	系数	标准差	系数	标准差
截距项	469.5***	19.85	468.7***	19.76	468.7***	19.76
年龄	-0.006***	0.00116	-0.00637***	0.00108	-0.00705***	0.00106
职业_蓝领	0.108***	0.03255	0.1546***	0.03248	0.1471***	0.03241
职业_企业家	0.09182	0.04933	0.1281**	0.04893	0.1193*	0.04888
职业_家政人员	-0.274***	0.0641	-0.2545***	0.06403	-0.2548***	0.06402
职业_管理员	0.1018**	0.0377	0.08484*	0.03776	0.08634*	0.03774
职业_退休	0.537***	0.05552	0.4847***	0.05608	0.4891***	0.05603
职业_自营	0.0198	0.05005	0.0514	0.04976	0.0471	0.04974
职业_服务业	0.08339*	0.03504	0.08892*	0.03511	0.08389*	0.03508
职业_学生	0.370***	0.09431	0.4537***	0.09359	0.4431***	0.09352
职业_技术人员	-0.04.35	0.02815	-0.06093*	0.02826	-0.06304*	0.02825
职业_无职业	-0.1457*	0.06353	-0.1413*	0.06357	-0.1548*	0.06347
婚姻_离婚	-0.05474	0.02941	——	——	——	——
婚姻_单身	0.0651**	0.02322	——	——	——	——
教育	0.013***	0.00319	0.0155***	0.00318	0.01676***	0.00315
违约_其他	-0.0649**	0.02265	-0.07384**	0.02269	——	——
房屋贷款	-0.03857*	0.01834	-0.03646*	0.01835	——	——
其他贷款	0.00809	0.02511	——	——	——	——
联络方式_手机	-0.484***	0.04047	-0.5574***	0.04099	-0.5578***	0.04099
月度联系总数	-0.001***	0.0000	-0.0005***	0.0000	-0.0005***	0.0000
联络次数	-0.022***	0.00363	-0.0221***	0.00364	-0.0221***	0.00364
上次联络次数	-0.366	0.2471	——	——	——	——
上次营销未联系	0.207	0.262	0.5792***	0.05526	0.5770***	0.05526
上次营销成功	0.997***	0.1897	0.7988***	0.1944	0.7924***	0.1944
员工变动率	2.141***	0.09674	2.148***	0.09624	2.142***	0.09623
消费价格指数	-1.229***	0.08477	-1.196***	0.08518	-1.195***	0.08508
消费者信心指数	-0.217***	0.00917	-0.2182***	0.00915	-0.2187***	0.00915
伦敦三月拆借利率	1.506***	0.07774	1.530***	0.07769	1.536***	0.07765
员工总数	-0.071***	0.00278	-0.07141***	0.002767	-0.07144***	0.002767
F1	0.4660143		0.4643245		0.4669018	
AUC	0.655		0.65		0.649	

从 BIC 准则下的回归结果可以对影响因素做出进一步的分析。

从客户个人信息来看，职业和受教育水平对营销是否成功的影响显著。职业因素方面，以行政人员(admin)为基准，退休人员(retire)和学生(student)是所有职业中最容易接受银行理财产品营销，而家政人员(housemaid)、无职业者(unemployed)和技术工作者(technician)是最容易拒绝银行营销的职业。教育水平方面，受教育程度越高，营销成功的可能性越大，但影响的程度没有职业的程度大。

从银行账户信息来看，BIC 准则下这类信息均被排除，但根据全模型和 AIC 准则下的模型，有过违约情况或房屋贷款的客户倾向于拒绝营销。

从营销联络情况的参数估计结果来看，座机联系会大大提高营销的成功率，而本次营销的集中联络次数及上月总共联络次数越多，营销的成功率会越低。

从营销联络次数信息来看，上次营销的结果对营销结果影响很大，与基准组上次营销失败对比，上次未进行营销和营销成功的客户在本次营销中成功的可能性很大，说明客户管理对银行很重要。

从经济指标来看，员工变动率、伦敦三月拆借利率和消费价格指数对营销成功影响最大，其中前两者正向影响成功率，消费者价格指数负向影响成功率。

3. 模型评价

在 30%的测试集上 Logit 回归的预测结果的混淆矩阵如下表所示。

表 3-3 Logit 模型预测结果			
		实际	
		0	1
预测	0	6662	1298
	1	2632	1721

从混淆矩阵可以计算得到，该模型的准确率有 68.08%，精确率有 39.5%，召回率有 57.0%，说明模型能够较好的找出更易于被营销成功的客户，与商业银行的目标一致，因此模型的预测效果较好，可以接受。

3.1.2 支持向量机模型

1. 模型介绍

支持向量机(Support Vector Machine)是由 Corinna Cortes 和 Vapnik 在 1995 年首次提出的一种算法[9]，在分类问题中，它可以处理线性可分和线性不可分的数据。对于线性不可分的数据[10]，通过非线性函数将数据映射更高空间中，在该维度寻找超平面来对数据进行分

类。考虑到现实数据通常为线性不可分数据和支持向量机在这种数据上的良好分类效果，本文采用高斯核函数的支持向量机对营销数据进行建模和预测。

2. 模型建立

由于支持向量机并不受不平衡样本影响，所以不需要对数据进行重新采样。对无序多分类建立哑变量，并且对所有数值型变量进行标准化，对处理过的数据建立高斯核函数的支持向量机。并且通过网格搜索寻找最优的模型参数 C 和 γ ， C 代表惩罚系数， γ 代表误差容忍度， C 较大的时候对于错误分类的容忍比较大， γ 决定了数据映射到新的特征空间后的分布。经过筛选，得到最优的参数 $C=100$ ， $\gamma=0.001$ 。

3. 模型评估

在 30% 的测试集上 SVM 预测结果的混淆矩阵如下表所示。从混淆矩阵可以计算得到，该模型的准确率为 90.1%，精确率有 73.8%，召回率只有 18.7%，综合得分 F1 有 0.298，说明该模型倾向于找出营销失败的客户。

表 3-4 SVM 模型预测结果

		实际	
		0	1
预测	0	10817	1106
	1	108	281

3.1.3 决策树模型

1. 模型介绍

决策树这一概念的产生源于归纳任务（the induction task）。在一项归纳任务中，我们有一组训练集，训练集中每一个样本有它的若干属性值以及一个类别值，目的是构建对属性值依次进行判断的一组规则，以此给出样本类别的归纳结果。这一判断规则可以形象化地以树形图表示，故称为决策树，判断过程从树的根部起始，自上而下依次递归，经一系列判断节点后到达叶子节点，给出归纳结果，故此类决策树得名 TDIDT（Top-Down Induction of Decision Trees[11]）。

TDIDT 起始于 CLS 算法（Hunt's Concept Learning System framework），由 Hunt、Marin 和 Stone 在 1966 年提出[15]。CLS 算法定义了分类的损失函数，其由两部分构成——测量样本某属性值的成本和分类错误的成本。而后，CLS 采用前瞻策略（Lookahead Strategy），在

每个节点上遍历一定深度决策树构成的整个空间，选取其中损失函数最小的一个，从而走向下一个节点。

CLS 算法固然可以挖掘出样本中的潜在模式，但成本驱动的前瞻策略所带来的运算量同样不可忽视，由此衍生出节点分裂的其他策略。1979 年，John Ross Quinlan 提出 ID3 (Iterative Dichotomiser 3) 算法[16][17]，采用信息增益最大准则，但该算法未考虑连续特征，且倾向选择子类较多的属性，具有较大的局限性。1984 年，Leo Breiman, Jerome Friedman, Richard Olshen 和 Charles Stone 提出 CART 算法 (Classification and Regression Tree) [13]，采用 Gini 不纯度最小准则。与此同时，CHAID (Chi-squared Automatic Interaction Detection) [18]、ID4 (Schlimmer & Fisher, 1986) [19]、ID5 (Utgoff, 1988) [20] 算法被陆续提出，在节点分裂策略和适用范围上各有不同。1993 年，Quinlan 本人对 ID3 算法加以改进，提出使用信息增益率最大作为准则的 C4.5 算法[11]，且以单点离散化思想解决连续变量的问题，在缺失值处理和剪枝技术等方面同样做出较大改进，成为目前使用最广泛的归纳算法之一。C5.0 算法是 C4.5 算法的商用版本，主要在执行效率和内存方面做出改进。

决策树以样本信息为基础构造分类和预测规则、挖掘内在规律模式，并不需要使用者对背景知识足够了解，且分类预测结果以树状图展示、便于理解和解释，因此在贷款风险评估、医学诊断、生产质量控制和天气预报等诸多领域得到广泛应用。

2. 模型建立

与之前一样，我们选取数据集的前 70%，并对营销成功个案进行重采样，作为训练集拟合模型；后 30% 作为测试集，用于测试模型的预测能力。我们使用 R 语言下的 `rpart` 程序包的 `rpart()` 函数构建决策树，其中方法选用 'gini'，即 CART 算法。我们首先使用 `rpart.control()`，将节点复杂度 (complex parameter) 设置为 0.01，对决策树进行预剪枝，而后使用 `prune()`，选择最小代价复杂度 (即使交叉验证误差最小的 `cp` 值)，自底向上对非叶节点进行考察，将不能带来泛化能力显著提升的非叶节点转为叶节点，从而对决策树进行后剪枝。决策树容易对训练集出现过拟合现象，为了提高模型在测试集上的泛化能力，剪枝是必要的操作。决策树的图形化展示如下。

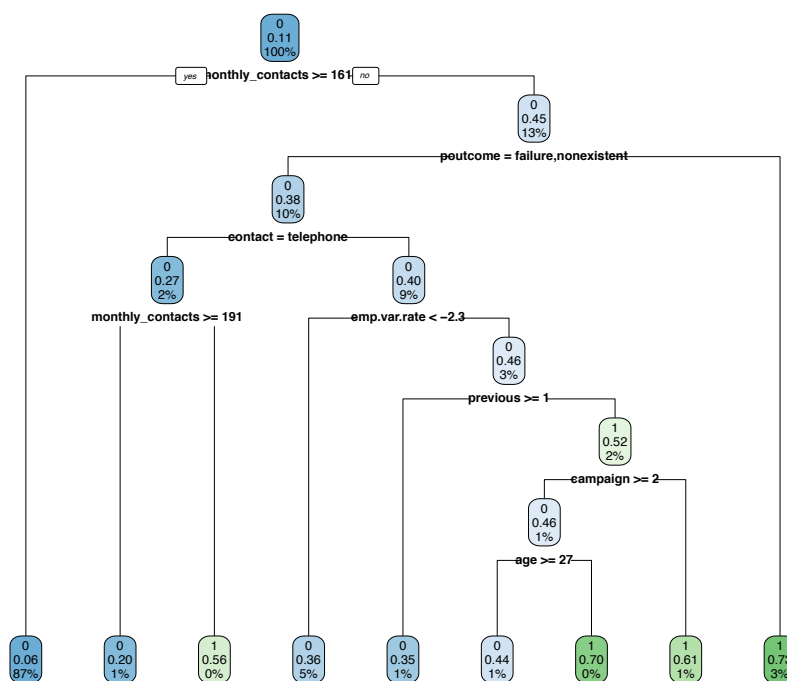


图 2-1 决策树可视化

由上图可见，决策树首先对当月总联系次数进行判断，当次数大于等于 1614 时，模型直接将营销结果预测为失败。否则，决策树再对上一次营销活动的结果（poutcome）进行判断，若上一次为成功，则预测此次也为成功，若上一次结果失败或未知，则要根据联系方式、员工周转率、先前的联系次数和年龄等属性依次判断，给出预测结果。观察决策树所呈现出的判断规则，我们不难发现以下规律：当月联系总次数较多时，营销结果更倾向于失败，这可能是因为业务员在每通电话上花费的时间和耐心更少；上一次营销活动成功的客户，这一次往往也更容易成功，他们应当成为银行密切关注和发展的客户群体；此次活动前的联系次数越多，结果往往越差，这可能是因为在被多次联系后，客户产生抵触情绪；对此次活动中联系次数的判断符合同样的规律；年轻客户（年龄小于 27 岁）更容易营销成功，这可能是因为年轻人有更多闲置资金、愿意接受新鲜事物等，而年龄稍大的人群可能对此类营销活动抱有更谨慎的态度。由此可见，决策树的分类结果易于观察和解释，且展现出的预测规则大都与我们的直观经验认识相符，银行可以据此调整目标客户群体和营销策略，这充分体现出决策树算法在此类分类任务中的优势。

3. 模型评估

在 30% 的测试集上决策树预测结果的混淆矩阵如下表所示。

表 3-5 决策树模型预测结果(临界值=0.5)

临界值=0.5	实际	
	0	1

预测	0	439	18
	1	8854	3001

可见，模型总体准确率仅 27.94%，我们进一步考察衡量模型预测效果的其他指标。模型精确率（查准效果）反映预测为成功的案例中实为成功的比例，为 25.31%；召回率（查全效果）反映实为成功的案例中成功预测的比例，为 99.40%。二者计算调和平均值，即为 $F1=40.35\%$ 。可见，模型对营销成功的情形过于灵敏，几乎可以全部识别，但与此同时却将大量营销失败的案例也预测为成功，大大损失了模型整体的准确率。银行若使用该模型，虽然能识别几乎全部潜在客户，但可能大大提高营销成本。故我们将预测的临界值调高至 0.75，重新预测得如下结果。可见，这时模型精确率有较大提升，模型整体准确率显著提升，但这是以召回率的降低为代价的。

表 3-6 决策树模型预测结果(临界值=0.75)

临界值=0.75		实际	
		0	1
预测	0	5264	524
	1	4029	2495

小二乘最小化的目标损失函数可能造成过拟合，子集选择方法虽然一定程度缓解，但是不能充分利用已有数据，因此在 3.2 中和 3.3 中，我们将使用两种压缩估计方法 Lasso 回归和 Ridge 回归，在上式基础上加入了正则化项，进一步解决目前遇到的过拟合与数据利用不充分的 trade-off 问题。

3.1.4 模型比较

本节采用了 Logit 模型、支持向量机和决策树三种模型对测试集进行预测。考虑到样本的分类不平衡，我们综合考虑准确率和召回率，使用两者的综合 $F_1 - measure$ 来衡量模型优劣。

表 3-7 模型效果比较

	accuracy	precision	recall	F1
Logit 模型	0.6808251	0.3953595	0.5700563	0.4669018
支持向量机	0.9009097	0.7378917	0.186734	0.2980437
决策树(0.5)	0.2794022	0.2531421	0.9940378	0.4035229
决策树(0.75)	0.6301982	0.3824341	0.8264326	0.5228964

根据上表可以看出，四项指标中，决策树的效果最好，其次逻辑回归的召回率较高，说明这两个模型都能最大程度的识别出可能营销成功的客户。因此在挖掘可以营销成功的用户时，应该首选决策树模型。而 SVM 模型的精确率达到 91%，召回率很低，适用于经费有限情况下，需要尽可能让每通电话都有效。Logit 模型的精确率和召回率为两者的折中，可

以需要根据成本收益的实际情况选择模型。此外，在研究营销策略时应该选择决策树模型和 Logit 模型，分析影响营销成功的因素。

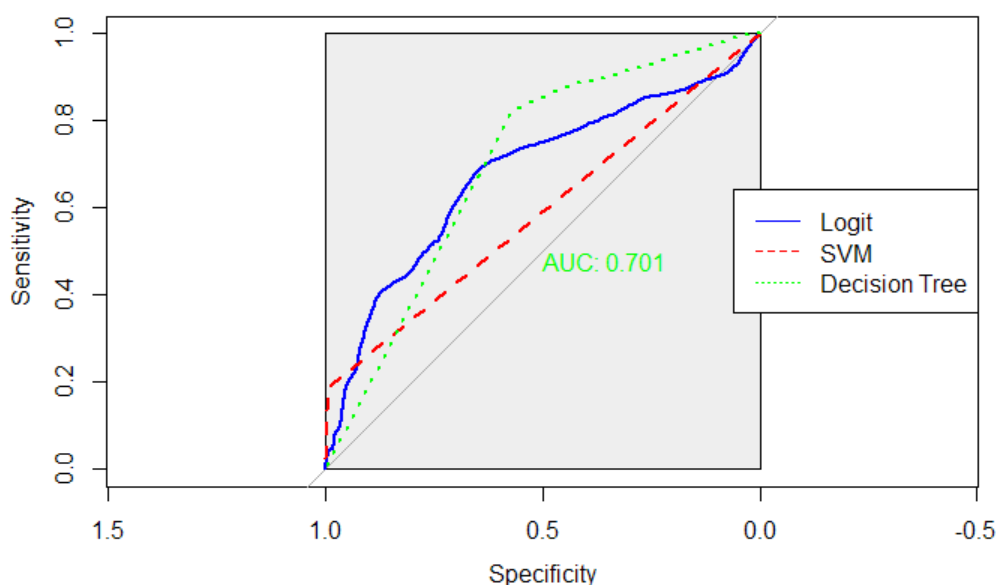


图 2-2 三个模型的 ROC 曲线

上图是三者的 ROC 曲线，ROC 以真阳性率作为纵坐标，假阳性率作为横坐标绘制而成，可以准确反映分类方法特异度和敏感度的关系，AUC(Area Under Curve)被定义为 ROC 曲线下与坐标轴围成的面积，AUC 越接近 1，分类方法的真实性越高，可以看到三个模型中决策树模型的 AUC 值最大，该决策树的 AUC 等于 $0.721 > 0.5$ ，表明模型并非随机猜测，而是具有一定的预测能力，但该预测能力尚可接受、并不十分显著。我们考虑产生这样结果的原因之一，是决策树算法自身的不稳定性，即决策树容易受数据的微小扰动影响，对此的解决方法之一是决策树的集成，即多次选取部分样本为训练集生成多棵决策树，采用投票等方式给出最终预测结果。

由 ROC 曲线图我们也可看出，在模型预测能力一定的情况下，模型灵敏度和特异度之间存在此消彼长的关系。银行的决策者究竟选择将潜在客户尽可能全部挖掘出来，还是选择保证每次营销的有效性，应考虑实际当中营销成本与收益间的平衡问题，从而给出利润最大化的最优方案。

3.2 聚类模型

3.2.1 聚类及用户画像综述

市场由有着不同兴趣、需求和背景的诸多主体组成，市场细分旨在将顾客依据其特征划分为若干顾客群，根据其不同需求进行差异化营销，从而提升顾客服务体验。聚类分析

(Clustering Analysis) 是一种无监督学习过程，其核心思想是“物以类聚、人以群分”，可以从无标记的数据中挖掘潜在结构，在市场细分方面有重要的应用价值。

聚类的概念最早由 Everitt 在 1974 年提出[22]，但迄今为止，学术界对聚类并没有公认的定义。聚类算法根据簇的不同生成方式，大体分为传统聚类和智能聚类。传统聚类包括划分聚类（Partitioning Clustering）与层次聚类（Hierarchical Clustering）两大类。划分聚类首先对数据集给出初始划分，而后根据一定的判定准则（如平方误差准则）不断迭代，直至满足结束准则（如聚类结果不再改变）为止。K-means 聚类是划分聚类的代表算法之一，由 J.B. MacQueen 在 1967 年提出[23]，后来许多学者对 K-means 算法做出大量改进，从不同角度克服其局限性，如 Pelleg 等提出的 X-means 算法[24]、Nguyen 等提出的 K-MODES 算法[25]等。层次聚类，又称系统聚类，顾名思义是对数据点进行逐层聚类，最终形成分层树形结构。基本的系统聚类包括聚合型和分裂型两种，由 Kaufman 和 Rousseeuw 提出。聚合型是指初始状态下每个数据点自成一类，而后逐次将距离最近的两类聚在一起，直至所有样本点聚为一类为止，分裂型则与之相反。分裂型算法的复杂度高、计算量大，实际并不常见。智能聚类则包括核聚类、人工神经网络聚类等算法，是当今机器学习和深度学习领域的理论前沿，限于篇幅，此处不再一一赘述。

两步聚类（Two-Step Clustering），亦称二阶聚类，是 Zhang 提出的 BIRCH（Balanced Iterative Reducing and Clustering using Hierarchies）[26]算法的改进版本。两步聚类将聚类过程分两步完成：第一步为预聚类，通过构建和修改聚类特征树，将样本点进行初步归类，形成诸多小的子簇；第二步为正式聚类，将预聚类所得的初步类别进行再聚类，将其逐个合并，确定最终聚类方案。在 SPSS 软件中，正式聚类采用层次聚类方法，且在每步根据 AIC 或 BIC 的大小以及类间最短距离的变化，自动确定最优聚类数。两步聚类可以处理同时包含定性与定量变量的数据，计算量小、运行速度快，符合我们所使用数据集的特征。

用户画像（persona）这一概念最早由交互设计之父 Alan Cooper 提出：“Personas are a concrete representation of target users” [27]，即根据一系列属性数据所抽象出的目标用户模型。精准的用户画像可以帮助公司精准营销、业务决策、为客户提供个性化服务。为了帮助银行锁定目标客户、提高营销效率，下面我们使用两步层次聚类对营销成功与失败的客户群体进行分类，给出颇具代表性的用户画像，以此为银行未来的营销方式提供指导，帮助其实现精准营销、降低成本。由于两步聚类中预聚类过程要求数据点完全随机排列，故我们将数据集顺序打乱，再进行两步聚类。

3.2.2 营销成功客户群体的聚类

我们首先对营销成功的客户个人信息进行两步聚类，挖掘该群体的特征。具有这些特征的个体倾向于办理存款业务，应是银行营销过程中密切关注和发展的客户。该群体被聚为 3 类，聚类轮廓系数约为 0.3，聚类效果尚好。

表 3-8 营销成功客户聚类结果

类别	年龄	职业	婚姻状况	教育程度	违约记录	住房贷款	个人贷款
1	32.26	行政	单身	高中/大学	无	有	无
2	42.49	行政/管理	已婚	大学	无	有/无	无
3	41.06	技术	已婚	专科	无	有	无

第一类客户画像：青年，行政岗位，单身，高中或大学学历，无违约记录，有住房贷款，无个人贷款；该人群大多尚未成家、收入稳定且有较大发展空间，有较多闲置资金可供投资，且对各种理财服务有较高热情，银行应充分利用手机、互联网等多种渠道开展营销，为其提供多样化的理财服务和产品，将其发展为稳定客户。

第二类客户画像：中年，行政或管理岗位，已婚，大学学历，无违约记录，有无房贷参半，无个人贷款；这类人群收入稳定、违约风险小，对未来的财务安排有一定规划，倾向于将闲置资金存入银行获取稳定收益。

第三类客户画像：中年，技术人员，已婚，受过专科教育，无违约记录，有房贷，无个人贷款。该人群虽受教育水平和收入水平不及前两类人群，但亦属于中产阶级，优于一般客户。

同时，2、3 类客户家中通常有小孩和老人需要抚养，银行可以向其提供教育或保险类金融投资产品，满足其特殊需求，提高客户粘性。

3.2.3 营销失败客户群体的聚类

下面，我们对营销失败且通话时长小于 10 秒的客户信息进行两步聚类，这类人群一般没有闲置资金用于存款，对营销内容不感兴趣，短时间内便挂断电话。银行在日后的营销过程中适当避免联系该类人群，可以提高营销效率。营销失败的客户群体同样聚为三类，聚类轮廓系数约 0.3，效果尚好。

表 3-9 营销失败客户聚类结果

类别	年龄	职业	婚姻状况	教育程度	违约记录	住房贷款	个人贷款
1	40.57	蓝领	已婚	基础教育	无	有	无
2	37.19	行政	单身	高中	无	有	无
3	37.38	技术	已婚	高中/专科	无	有/无	无

第一类客户画像：中年，蓝领，已婚，仅受过基础教育，没有违约记录，大多数有住房

贷款，无个人贷款；这类人群受教育水平低、主要从事技术含量较低的体力劳动，薪酬较低、需要偿还住房贷款，一般没有闲置资金用于存款；

第二类客户画像：中青年，行政，单身，高中学历，无违约记录，有房贷，无个人贷款；该类人群受教育程度较低，可能在行政单位从事比较基础的工作岗位，同时有住房贷款需要偿还，亦没有闲置资金用于存款。

第三类客户画像：中青年，技术人员，已婚，受过高中或专科教育，没有违约记录，房贷有无参半，没有个人贷款。该类人群与营销成功的第三类人群相比，受教育水平偏低，因此工资待遇可能也更差。

经过以上分析，我们已利用聚类算法挖掘出若干客户群体画像，结合经验认识给出解释，对银行的不同客户群体有了粗略认识，由此初步验证了聚类算法在市场细分领域的应用价值。然而，由于部分群体（如离异群体、学生群体、服务行业的从业人员）在数据集中出现频次过小，其特征并未得到充分体现。后续，我们可以利用更大容量的客户信息，或添加其他变量（如账户余额、家庭成员、工资薪酬等），对客户特征进行更细腻的观察和刻画，从而帮助银行更精准地锁定目标客户群体。

4. 结论

本文首先描述了整体的数据准备方案，包含大量基于描述性分析的数据处理操作，如提取每月电话营销数的关键指标，通过分组小提琴图、mosaic 图、云雨图等对多个离散型特征进行交叉可视化，对进一步特征的删除、合并、转义操作提供了可视化支持。基于特征筛选，我们实现了多多数缺失值的处理，并最后对剩余的观测使用了随机森林、多重插补等方式处理了缺失，保证了进入模型数据质量的最大化。

随后，我们采用 Logit 回归、支持向量机和决策树三种模型进行银行电话营销成功与失败的建模、预测与评估，并结合实际对有关结果进行解读。我们发现，高召回率的决策树模型适用于需要尽可能多的获得成功营销客户的情况，高准确率的 SVM 模型适用于经费有限的情况，而 Logit 模型可以在这两者中进行权衡。此外，我们在题目要求的基础上进行创新性延伸，利用两步层次聚类算法对营销成功和失败的客户群体分别进行聚类 and 比较，挖掘出两类群体的若干典型特征，为银行进一步精准营销客户管理提供有价值的信息。

虽然我们仅以 Logit 回归、支持向量机和决策树为例，建立银行电话营销成功与失败的预测模型，但成功从结果中挖掘出部分一般规律，以便为银行改善营销策略提供指导性建议，具备比较显著的现实意义。未来，我们可以对上述算法进一步优化，提高其预测性能，同时也可以引入更多分类模型构成集成分类器，在不同模型间取长补短，得到更加准确可靠的预测结果。

由于时间、知识水平以及篇幅所限，本文一定有诸多不完整或不准确之处，恳请老师批评指正，谢谢！

参考文献

- [1] 李冬馥. 商业银行市场营销研究理论综述[J]. 牡丹江大学学报, 2010(03):115-117.
- [2] 宋建华. 商业银行电话营销[J]. 金融论坛.2015(10).
- [3] 戴轶, 姚兵. 银行关系营销与客户关系管理[J]. 行政事业资产与财务, 2014(3): 35-37.
- [4] Elsalamony, H. A. (2014). Bank direct marketing analysis of data mining techniques. *International Journal of Computer Applications*, 85(7), 12-22.
- [5] 高丽. (0). 银行电话营销成功之路的分析与预测. (Doctoral dissertation).
- [6] Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62, 22-31.
- [7] 陈锟, 王晓红, & 朱敏. (2006). Logit 模型在市场营销定量研究中的应用——消费者对家电品牌选择的实证分析. *商业研究*, 000(024), 195-199.
- [8] 李萌. (2005). Logit 模型在商业银行信用风险评估中的应用研究. *管理科学*, 18(2), 33-38.
- [9] Cortes, C., & Vapnik, V. (1995). Support vector machine. *Machine learning*, 20(3), 273-297.
- [10] Krishna, C. L., & Reddy, P. V. S. (2019). An Efficient Deep Neural Network Classifier in Banking System. *International Journal of Research and Analytical Reviews*, 6(2), 511-517.
- [11] Quinlan J R . Induction on decision tree[J]. *Machine Learning*, 1986, 1.
- [12] Quinlan J R . C4.5: Programs for Machine Learning[M]. 1993.
- [13] Friedman. Classification and Regression Trees[M]// *Classification and regression trees*.
- [14] Breiman L I , Friedman J H , Olshen R A , et al. Classification and Regression Trees. Wadsworth.[J]. *Biometrics*, 1984, 40(3):358.
- [15] Hunt E B , Marin J , Stone P J . Experiments in induction.[J]. *The American Journal of Psychology*, 1966, 80(4).
- [16] Quinlan, J.R. *Discovering Rules by Induction from Large Collections of Examples*. Expert Systems in the Micro Electronic Age. Edinburgh University Press, 1979.
- [17] Quinlan J R . LEARNING EFFICIENT CLASSIFICATION PROCEDURES AND THEIR APPLICATION TO CHESS END GAMES[J]. *Machine Learning*, 1983:463-482.
- [18] Gordon V. K. (1980).An Exploratory Technique for Investigating Large Quantities of Categorical Data, *Applied Statistics*. 29(2): 119–127.
- [19] Schlimmer J C , Fisher D . A Case Study of Incremental Concept Induction[C]// *National*

-
- Conference on Artificial Intelligence Philadelphia. DBLP, 1986.
- [20] Utgoff P E . ID5: An Incremental ID3.[C]// International Conference on Machine Learning. 1988.
- [21] Ball G, Hall D. ISODATA, a novel method of data analysis and pattern classification[M]. California: Technical rept. NTISAD 699616. Stanford Research Institute, 1965.
- [22] Everitt, B. (1974) Cluster Analysis. Heinemann Educational Books Ltd., London.
- [23] Macqueen J . Some methods for classification and analysis of multivariate observations[C]// Proc. Fifth Berkeley Symposium on Math. Stat. and Prob. 1967.
- [24] Pelleg D . X-means : Extending K-means with Efficient Estimation of the Number of Clusters[C]// Proc. 17th International Conference on Machine Learning, 2000. 2000.
- [25] Nguyen H H . Privacy-Preserving Mechanisms for k-Modes Clustering[J]. Computers & Security, 2018, 78(sep.):60-75.
- [26] Zhang T , Ramakrishnan R , Livny M . BIRCH: An Efficient Data Clustering Method for Very Large[J]. acm sigmod record, 1996, 25(2):103-114.
- [27] Cooper A , Reimann R , Cronin D , et al. About Face: The Essentials of Interaction Design[J]. TORI, R.; BERNARDES, J.L.; BIANCHINI, R.; JACOBBER, E.C. 2003a. A Practical Study on the Usage of a Commercial Game Engine for the Development of Educational Games In: Proceedings of SBGames WJogos, 2007, 18(1):1.
- [28] Smith, Wendell R . Product Differentiation and Market Segmentation as Alterative Marketing Strategies [J]. Journal of Marketing, 1956: 3-8.
- [29] H.W.Shin, S.Y.Sohn.Segmentation of stock trading customers according to potential value[M] . Expert Systems with Applications, 2004 (27): 27-33.