# 银行客户长期存款产品购买情况预测

唐露函（202011011028）

北京师范大学统计学院

**摘要:**本报告基于Banking Dataset Classification数据集，以年龄、工作、婚姻、教育、默认、住房、贷款、接触、月、一天的一周、期间、活动、结果13个变量为自变量，对银行客户是否会（是/否）订购该银行的定期存款产品进行预测，属于分类问题。首先，针对数据集进行了预先处理与初步分析，随后运用Logistic 回归、线性判别分析、KNN分类器与决策树4种不同的模型建模预测。其中采用CV准则对KNN模型的K值进行了选择，并对决策树进行了剪枝处理。最后，所有模型以精确率、召回率、特异度、ROC曲线和AUC值为评价指标进行了比较与分析。结果显示，Logistic 回归、线性判别分析与决策树模型的AUC值分别为：0.8547613、0.8589598、0.8425568，有很好的预测效果，且在精确率、召回率方面,这三个模型均高于91%,预测效果显著优于KNN分类器模型。本报告通过比较、评估不同模型效果，提高了预测结果的精准度，同时也为银行利用模型解决客户购买问题提供了不同的决策思路。

**关键词:** 分类；Logistic；LDA；KNN；决策树

## 一、 引 言

### （一） 背景引入

随着金融业多元化经营法律环境和政策条件的不断改善,国内外各商业银行或通过成立控股子公司、或通过内部创新等纷纷开展多元化业务,推进综合化经营战略,以创造新的利润增长点。大多数银行都将多元化经营作为实现战略转型、提升综合竞争力的必要手段,积极探索在证券、保险、基金、信托等相关金融领域的有关业务。这些业务收入都是银行运作盈利的重要手段。2008 年 9 月，雷曼兄弟公司在美联储拒绝资金援助后申请破产保护，标志着由“次级贷”引发的金融危机在全球范围内全面爆发。起源于美国的这场危机产生的连锁反应致使全球信贷市场陷入混乱，对国际金融秩序造成了极大的冲击和破坏，同时也使得金融市场产生了强烈的信贷紧缩效应。（王亮亮[1]）

在这场金融危机中，一家葡萄牙银行的收入下降了，银行管理人员想知道他们该采取什么行动来应对。经过调查，他们发现根本原因是客户的长期存款投资不足。因此，该银行想要确定那些有更高机会购买长期存款的现有客户，并将营销工作集中在这些客户身上。数据

与葡萄牙银行机构的直接营销活动有关。市场营销活动以电话为基础。通常，银行工作人员需要致电访问客户是否要订购银行定期存款（“是”）还是不订购（“否”）。

**（二） 研究目的及数据介绍**

本分类研究的目的是预测该银行的客户是否会（是/否）订购定期存款。

数据平台 Kaggle 中的 Banking Dataset Classification 收集了从 2008 年 5 月到 2010 年 11 月期间该银行客户的相关数据，包括 15 个自变量$X$以及 1 个因变量$y$，变量及其描述如下表所示。

表 1 变量及其描述

| 变量 | 类型 | 描述 |
|---|---|---|
| $x_1$ 年龄 age | 数字 | 一个人的年龄 |
| $x_2$ 工作 job | 类别 | 工作类型（管理员，蓝领，企业家，女佣，管理，退休，自雇，服务，学生，技术员，失业，未知） |
| $x_3$ 婚姻 marriage | 类别 | 婚姻状况（离婚，已婚，单身，未知） |
| $x_4$教育 education | 类别 | （高中，文盲，专业课程，大学学位，未知） |
| $x_5$ 默认 default | 类别 | 有违约信用吗？（“否”, “是”, “未知”） |
| $x_6$ 住房 housing | 类别 | 有住房贷款吗？（“否”, “是”, “未知”） |
| $x_7$ 贷款 loan | 类别 | 有个人贷款吗？（“否”, “是”, “未知”） |
| $x_8$ 接触 contact | 类别 | 联系人通信类型（“蜂窝”, “电话”） |
| $x_9$ 月 month | 类别 | 一年中的最后一个联系月份（1 月到 12 月） |
| $x_{10}$ 一天的一周 day_of_week | 类别<br>数字 | 一周中的最后一个联系日（周一到周五） |
| $x_{11}$期间 duration | 数字 | 上一次联系时间，以秒为单位 |
| $x_{12}$活动 campaign | 数字 | 在此广告系列期间以及为此客户执行的联系数量（包括最后一次联系） |
| $x_{13}$ 天数 pdays | 数字 | 自上次广告系列最后一次联系客户以来经过的天数（999 表示以前未联系过客户） |
| $x_{14}$以前 previous | 数字 | 此广告系列之前以及为此客户执行的联系数量 |
| $x_{15}$结果 poutcome | 类别 | 先前营销活动的结果（“失 |

| | | | |
|---|---|---|---|
| | | | 败”，“不存在”，“成功”） |
| $y$ | 是否订购产品 | 类别 | 客户是否已订阅定期存款？（“是”，“否”） |

数据来源：Banking Dataset Classification

# 二、 建模和估计方法

## （一） 数据预处理

### 1. 相关系数

Spearman 相关系数：

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

其中，$d^i$为$(x_i, x_j)$秩次的差值，n 为化学成分变量中数据的个数。

$\rho > 0$则两个变量呈正相关关系；$\rho < 0$则两个变量呈负相关关系。在热力图中，ρ的绝对值|ρ|越接近 1，即颜色越深，表明正/反相关关系越强。|ρ|越接近 0，即颜色越浅，则表示两个变量间相关关系越弱。

## （二） 模型介绍

### 1. Logistic 回归模型

根据因变量$y$客户是否订购产品是一个二分类变量(“否”或者“是”,记为“0”与“1”),而不是一个连续变量,所以对于二分类因变量的分析需要使用非线性函数（于立勇等人[2]）。

研究二元响应变量$y$和 13 维预测变量$X$之间的关系，建立 logistic 回归模型：

$$P(y=1|X) = \frac{e^{\beta_0+\beta^T X}}{1+e^{\beta_0+\beta^T X}} =: P(X)$$

其中，

$y = 1$或$0$，为因变量

$X = (x_1, x_2, \dots, x_{13})^T$是 13 维的自变量

$\beta_0$和$\beta^T = (\beta_1, \dots, \beta_{13})$是未知的回归系数

$P(X)$是 logistic 函数

由$P(X)$可得似然函数：

$$l(\beta_0, \beta) = \prod_{i=1}^{n} P(x_i)^{y_i} (1 - P(x_i))^{1-y_i}$$

对数似然函数为：

$$L(\beta_0,\beta)=\sum_{i=1}^{n}\{y_i(\beta_0+\beta^T x_i)-\log(1+e^{\beta_0+\beta^T x_i})\}$$

关于$\beta$和$\beta_0$极大化以上对数似然函数，可得到回归系数的极大似然估计为：

$$(\widehat{\beta_0},\hat{\beta})=argmax_{\beta_0,\beta}L(\beta_0,\beta)$$

运用 Newton-Raphon 迭代方法进行求解：

$$\theta^{(k+1)}=\theta^{(k)}-(\frac{\partial^2 l(\theta)}{\partial\theta\partial\theta^T}|_{\theta=\theta^{(k)}})^{-1}\frac{\partial l(\theta)}{\partial\theta}|_{\theta=\theta^{(k)}}$$

其中，

$\theta=(\beta_0,\beta)^T$

$\theta^{(k)}$表示第 k 步的$\theta$的估计值

对于给定的$\in$（可取$10^{-5}$），当$\|\theta^{(k+1)}-\theta^{(k)}\|^2<\in$时，停止迭代。

**2. 线性判别分析模型（LDA）**

线性判别分析模型，Logistic 回归，建立 $P(y=k|X=x)$模型，即给定预测变量$X$条件下，建立响应变量$y$的条件分布模型。

根据数据分布的正态性，运用 Bayes 定理估计$P(y=k|X=x)$，来进行分类。

$$f(x)=\frac{1}{(2\pi)^{p/2}|\sum|^{1/2}}\exp(-\frac{1}{2}(x-\mu)^T\sum^{-1}(x-\mu))$$

其中，

$f(x)$是多元正态密度函数

$X=(x_1,x_2,\dots,x_{13})^T$是 13 维的，服从一个均值不同，协方差矩阵相同的多元正态分布

$\mu=(\mu_1,\dots,\mu_p)^T$，$\sum$ 是协方差矩阵

Bayes 分类器将$X=x$ 分到使得下式最大的一类：

$$\delta_k(x)=x^T\widehat{\sum}^{-1}\widehat{\mu_k}-\frac{1}{2}\widehat{\mu_k}^T\widehat{\sum}^{-1}\widehat{\mu_k}+log\widehat{\pi_k}$$

在本模型中，y 的预测结果共有两类，故 k 取值为 0 或 1

**3. KNN 分类器模型**

KNN 算法的基本思路（张宁等人[3]）是: 以所有已知类别的样本作为参照，考虑在训练集中与该未知样本距离最近(最相似)的 K 个已知样本,根据这 K 篇文本所属的类别判定新文本所属的类别, 以少数服从多数的投票法则（majority-voting），将未知样本与 K 个最邻近样本中所属类别占比较多的归为一类。

KNN 方法运用贝叶斯准则将观测值$x_0$分到概率最大的类中。

条件概率$P(Y=j|X=x_0)$，Y 的可能取值为 0 或 1，如果$P(Y=j|X=x_0)>0.5$，贝叶斯分类器将该观测的类别预测为 1。

给定 K 值和预测点$x_0$，KNN 分类器首先确定 K 个最靠近$x_0$的训练观测，记为$N_0$。然后对每个类别 j 分别用$N_0$中的点估计一个分值作为条件概率的估计（James, G.等人 [4]），这个值等于 j：

$$P(Y=j|X=x_0)=\frac{1}{K}\sum_{i\in N_0} I(y_i=j)$$

对于 K 值的选择：取小的 K 值提供了最灵活的拟合，但将导致偏差变小和方差变大；取大的 K 值提供的拟合更平滑，但将导致偏差变大和方差变小。本报告将尝试多个 K 来查看预测效果。

**4. 决策树模型**

本次研究的目标变量是定性变量，故选择分类决策树模型（吉根林等人[5]）。分类决策树模型建立过程：

采用递归二叉分割方法，以分类错误率作为分类点的准则。其中分类错误率的表达式为：

$$E=1-\max_k(\hat{p}_{mk})$$

其中，$\hat{p}_{mk}$代表第 m 个区域的训练集中第 k 类所占比例。

分类树偏差可以说明树拟合数据效果，偏差小说明拟合效果好，其表达式为：

$$-2\sum_m\sum_k n_{mk}log\hat{p}_{mk}$$

其中，$n_{mk}$是第 m 个终端节点处属于第 k 类观测值的个数。

**（三） 分类模型的评价指标**

**1. 混淆矩阵**（邹洪侠等人[6]）

通常以关注的类别为正类，其他类为负类。分类算法在测试集上的预测或正确或者不正确，共有 4 种可能的情况。可排列在如下所谓混淆矩阵中：

表 2 混淆矩阵（Confusion Matrix）

| | | 预测值 | |
|---|---|---|---|
| | | 正类 | 负类 |
| 真实值 | 正类 | TP（True Positive 真阳性） | FN（False Positive 假阴性） |
| | 负类 | FP（False Positive 假阳性） | TN（True Negative 真阴性） |

第一个字母表示预测正确与否，T(True)表示预测正确，F(False)表示预测错误；第二个

字母表示分类算法的预测结果，P(Positive)表示预测为正类，N(Negative)表示预测为负类。

基于混淆矩阵，有以下几个指标：

**2. 准确率（Accuracy）**

准确率是分类算法正确分类的样本数和总样本数之比。值得注意的是： 在类别样本不均衡的情况下，占比大的类别往往会成为影响准确率的最主要因素，此时的准确率并不能很好地反映模型的整体情况。其定义如下：

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

**3. 精确率（Precision）**

精确率是分类算法预测的正样本中预测正确的比例，取值范围为[0,1]，取值越大，模型预测能力越好，其定义为：

$$Precision = \frac{TP}{TP + FP}$$

**4. 召回率（Recall）**

召回率是分类算法所预测正确的正样本占所有正样本的比例，取值范围为0,1，取值越大，模型预测能力越好。其定义为：

$$Recall = \frac{TP}{TP + FN}$$

**5. ROC 曲线**

在分类任务中，测试部分通常是获得一个概率表示当前样本属于正例的概率， 我们往往会采取一个阈值，概率大于该阈值的为正例，小于该阈值的为负例。如果我们减小这个阈值，那么会有更多的样本被识别为正类，这会提高正类的识别率，但同时会降低负类的识别率。为了形象地描述上述这种变化，引入 ROC 曲线来评价一个分类算法的好坏。ROC 曲线全称为"受试者操作特性曲线" (Receiver Operating Characteristic Curve)。

ROC 曲线的纵轴是不同阈值水平下的真阳性率（TPR，TruePositive Rate），即在所有的正样本中，分类算法预测正确的比例：

$$TPR = \frac{TP}{TP + FN}$$

TPR 实际上就是 Recall 召回率，也被称为灵敏度(Sensitivity)。

横轴

ROC 曲线的横轴则是不同阈值水平下的假阳性率，即在所有的负样本中，分类器预测错误的比例：

$$FPR = \frac{FP}{FP + TN}$$

1-FPR 也被称为特异度(Specificity)。

**6. AUC**

判断不同模型算法的分类效果，较为合理的判据是比较ROC 曲线下的面积，即 AUC(Area Under ROC Curve)。AUC 的值越大表示预测效果越好。实际上，若 AUC=1，表示完美分类算法； 若 0.5<AUC<1，表示分类算法优于随机猜测。选取适当的阈值，有一定的预测价值。若 AUC=0.5，则和随机猜测效果一样，模型没有预测价值；AUC< 0.5，则比随机猜测效果还差。

## 三、 数值模拟与实例分析

### （一） 数据预处理

本报告初步拟定通过$x_1, x_2, \ldots, x_{15}$来预测$y$，由于$x_i$中包含部分非数值型变量，因此在建立模型之前，需要先进行数据的预处理，并进行简单分析。

**1. 缺失值处理**

数据集中含有“unknown”字段即表示数据缺失。由于总数据集数据量庞大，共有 32950 条，因此，删除含有“unknown”的数据。

**2. 变量去除**

由于$x_{13}$（pdays）和$x_{14}$（previous）主要只包含一个相同的数值，二者的方差很小，技术上对预测没有帮助，因此去掉这两个变量。

将剩余变量重新命名为$x_{1,\ldots,}x_{13}, y$,共 13 个自变量和 1 个因变量。

**3. 离群值处理**

画出变量的直方图，发现$x_1$（age）、$x_{11}$（duration）、$x_{12}$（campaign）三个变量中含有异常值。因此，将三个变量中的离群点替换为变量各自的上下边界。

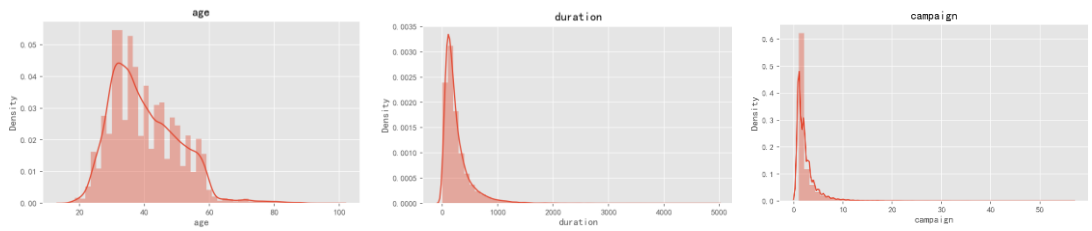


图 1 age、duration、campaign 直方图

**4. 编码分类特征**

根据后续的建模需要，需要将分类的变量进行编码转化为数值。转化结果见附录表 1 至

表 10。值得注意的是，将目标变量 y 编码为“是”=1，“否”=0。

**5. 划分测试集与训练集**

划分训练集与测试集（7：3），训练集数据用于拟合模型，测试集数据用于计算模型的准确率（Accuracy）、精确率（Precision）、召回率（Recall）以及绘制 ROC 曲线、计算 AUC。

**（二） 数据初步分析**

**1. 数据平衡性**

在数据集中，有约 88.73%的客户没有订阅定期存款，即$y = 0$；有 11.27%的客户订阅了定期存款，即$y = 1$。目标变量$y$的类分布为约 89：11，表示数据集不平衡。

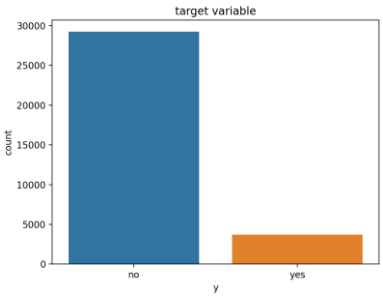


图 2 目标变量类分布

查看$x_i$的具体数量情况，可得出以下初步结论：客户职业前三多的是行政，蓝领工作和技术人员；在订阅定期存款的客户中，从事行政工作的客户占大多数；大量客户都结婚了；大多数客户没有违约信用；过去的许多客户都申请了住房贷款，但很少有人申请过个人贷款；手机似乎是接触客户最青睐的方法；银行在 5 月份已经联系了许多客户。

**2. 相关性分析**

首先进行正态性检验，得到结果：$x_{1,\ldots,}x_{13}$的显著性都低于 5%，拒绝原假设，不服从正态分布。因此选择用 Spearman 相关系数来探究相关性。

表 3 正态性检验

| | 柯尔莫戈洛夫-斯米诺夫(V)a 里利氏显著性修正 | | |
|---|---|---|---|
| | 统计 | 自由度 | 显著性 |
| age | .095 | 32950 | .000 |
| job | .257 | 32950 | .000 |
| marital | .330 | 32950 | .000 |
| education | .203 | 32950 | .000 |
| default | .504 | 32950 | .000 |
| housing | .366 | 32950 | .000 |
| loan | .512 | 32950 | .000 |
| contact | .411 | 32950 | .000 |

| | | | |
|---|---|---|---|
| month | .242 | 32950 | .000 |
| day_of_week | .161 | 32950 | .000 |
| duration | .132 | 32950 | .000 |
| campaign | .255 | 32950 | .000 |
| poutcome | .473 | 32950 | .000 |

画出自变量$x_i$的相关系数热力图，没有高度相关的特征。

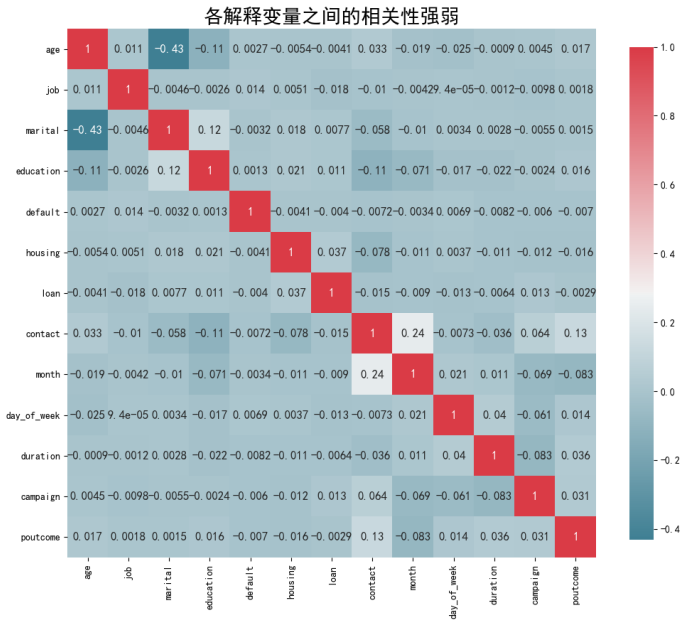


图 3 相关性分析

基于以上的数据处理与分析，为预测客户订购产品的情况（“是”或“否”），下面拟定用 Logistic 回归模型、线性判别分析模型、KNN 分类器模型和决策树模型共 4 种不同的模型来进行预测研究，并对各模型预测效果进行对比，选择出最优的预测模型。

### （三） 建模预测

#### 1. Logistic 回归模型

利用 R 语言对预处理好的数据建立 Logistic 回归模型，结果如下：

表 4 Logistic 回归结果

| | Estimate | Std. Error | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|
| (Intercept) | -2.550e-01 | 1.390e-02 | -18.338 | < 2e-16 | *** |
| age | 1.441e-03 | 2.024e-04 | 7.117 | 1.13e-12 | *** |
| job | 1.436e-03 | 5.332e-04 | 2.693 | 0.00708 | ** |
| marital | 2.731e-02 | 3.408e-03 | 8.014 | 1.17e-15 | *** |
| education | 8.080e-03 | 9.259e-04 | 8.727 | < 2e-16 | *** |
| default | -3.501e-02 | 1.647e-01 | -0.213 | 0.83168 | |
| housing | 4.520e-03 | 3.786e-03 | 1.194 | 0.23258 | |

| | | | | | |
|---|---|---|---|---|---|
| loan | -2.328e-03 | 5.234e-03 | -0.445 | 0.65652 | |
| contact | -9.361e-02 | 4.149e-03 | -22.560 | < 2e-16 | *** |
| month | 5.504e-03 | 8.512e-04 | 6.466 | 1.03e-10 | *** |
| day_of_week | -1.511e-03 | 1.344e-03 | -1.125 | 0.26076 | |
| duration | 7.053e-04 | 1.066e-05 | 66.161 | < 2e-16 | *** |
| campaign | -5.851e-03 | 1.231e-03 | -4.753 | 2.01e-06 | *** |
| poutcome | 1.120e-01 | 5.209e-03 | 21.503 | < 2e-16 | *** |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

根据上表，可以看出大部分变量都是显著的，但仍有 4 个变量：$x_5$、$x_6$、$x_7$、$x_{10}$是不显著的。下面将四个不显著的变量删去，重新进行拟合，对两个回归方程的结果进行比较分析。

（1）利用所有解释变量$x_{1,\ldots,}x_{13}$拟合的 logistic 回归模型

混淆矩阵为：

表 5 混淆矩阵

| | 0 | 1 |
|---|---|---|
| 0 | 5923 | 542 |
| 1 | 172 | 178 |

准确率（Accuracy）= 0.8953693

精确率（Precision）= 0.9717801

召回率（Recall）= 0.916164

灵敏度(Sensitivity) = 0.916164

特异度(Specificity) = 0.5208914

ROC 曲线如下图：

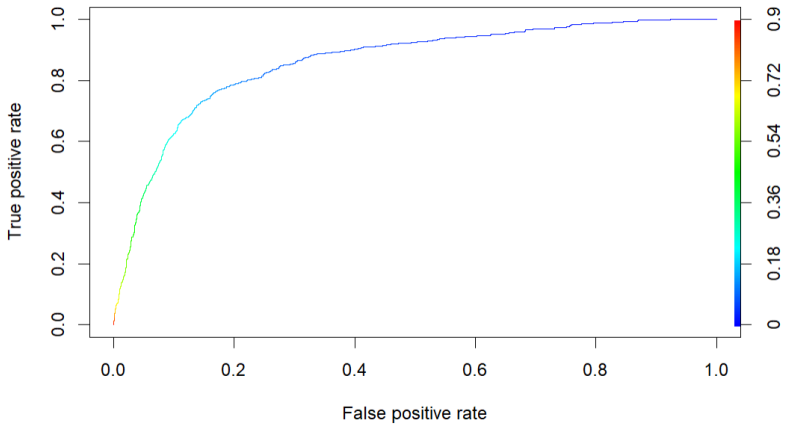


图 4 ROC 曲线

AUC = 0.8589598

（2）删除$x_5$、$x_6$、$x_7$、$x_{10}$拟合的 logistic 回归模型

混淆矩阵为：

表 6 混淆矩阵

| | 0 | 1 |
|---|---|---|
| 0 | 5933 | 548 |
| 1 | 162 | 181 |

准确率（Accuracy）= 0.8959555

精确率（Precision）= 0.9734208

召回率（Recall）= 0.9154451

灵敏度(Sensitivity) = 0.9154451

特异度(Specificity) = 0.5276968

ROC 曲如下图：

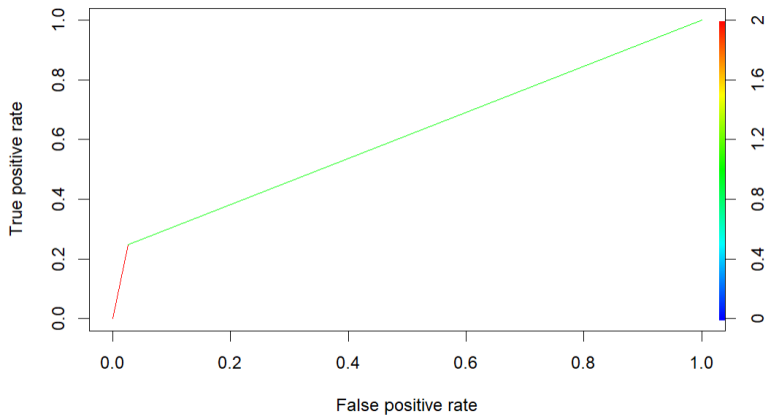


图 5 ROC 曲线

AUC = 0.6108531

由于 AUC 值的差异，选择 AUC 值更大的、不删除变量的模型，下面写出 Logistic 回归模型估计方程：

$$P(y=1|X)=\frac{e^{f(x)}}{1+e^{f(x)}}=:P(X)$$

$$f(x)=-0.2549578096+0.0014408498x_1+0.0014362939x_2+0.0273111423x_3+0.0080802643x_4-0.0350109019x_5+0.0045200607x_6-0.0023277120x_7-0.0936052216x_8+0.0055036132x_9-0.0015110232x_{10}+0.0007052944x_{11}-0.0058511946x_{12}+0.1120176261x_{13}$$

## 2. 线性判别分析模型（LDA）

利用 R 对训练集进行 LDA 模型拟合，并用测试集中的数据进行检验，结果如下：


图 6 测试集真实值

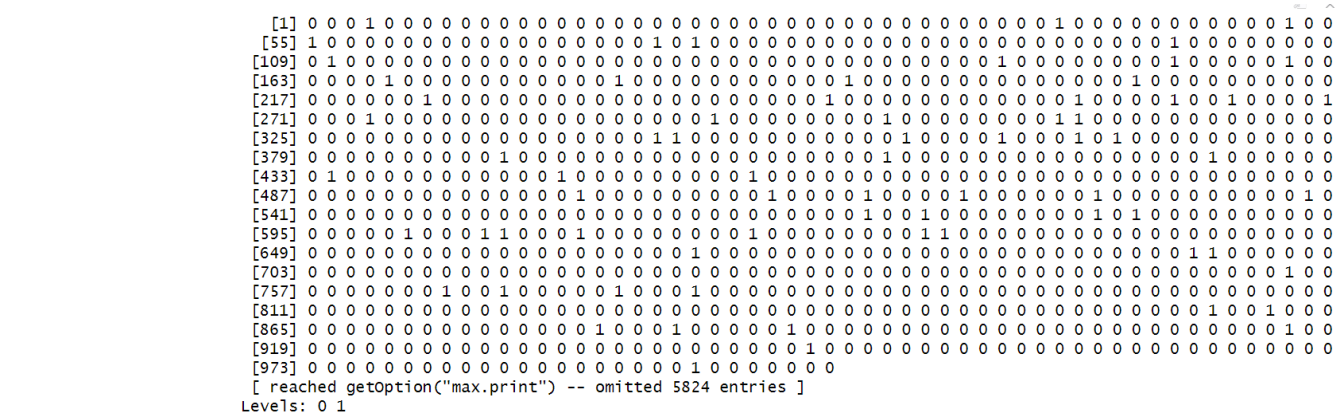


图 7 测试集预测值

混淆矩阵为：

表 7 混淆矩阵

| | 0 | 1 |
|---|---|---|
| 0 | 5800 | 438 |
| 1 | 295 | 291 |

准确率（Accuracy）= 0.892585

精确率（Precision）= 0.9515997

召回率（Recall）= 0.9297852

灵敏度(Sensitivity) = 0.9297852

特异度(Specificity) = 0.496587

ROC 曲如下图：

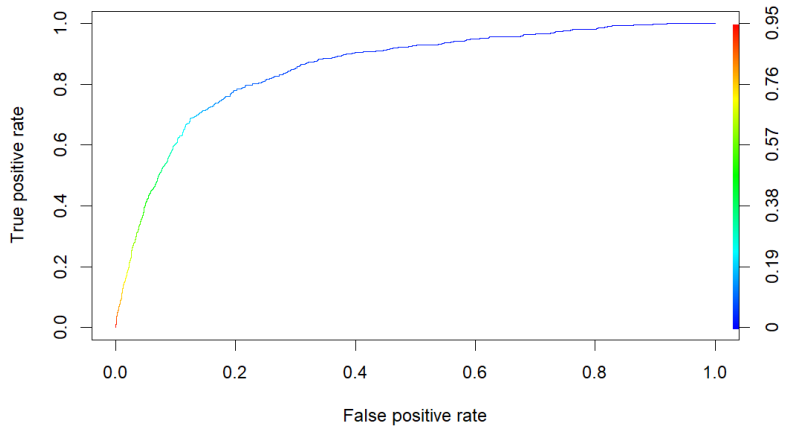


图 8 ROC 曲线

AUC = 0.8547613

### 3. KNN 分类器模型

利用 R 对 KNN 模型进行构建，选取不同的 K，得到了对应的准确率和 AUC 值。

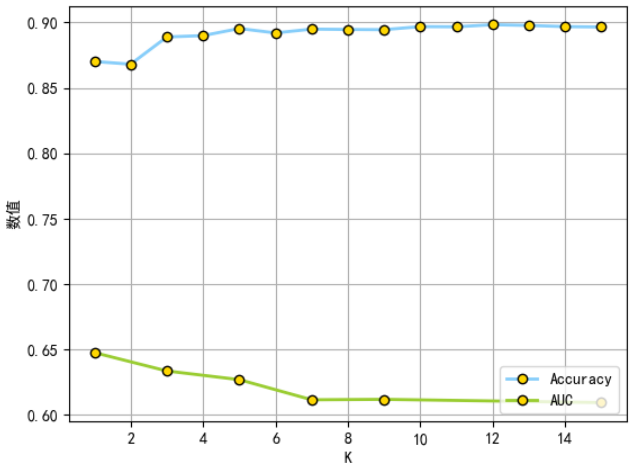


图 9 K 对应的准确率和 AUC

当 K=1 时，准确率约为 0.87，与 K=3 时的准确率相差不大。但 K=1 时 AUC 值最大，因此选择 K=1 来进行模型预测。

混淆矩阵为：

表 8 混淆矩阵

| | 0 | 1 |
|---|---|---|
| 0 | 5678 | 417 |
| 1 | 466 | 263 |

准确率（Accuracy）= 0.8706038

精确率（Precision）= 0.9241536

召回率（Recall）= 0.9315833

灵敏度(Sensitivity) = 0.9315833

特异度(Specificity) = 0.3607682

ROC 曲如下图：

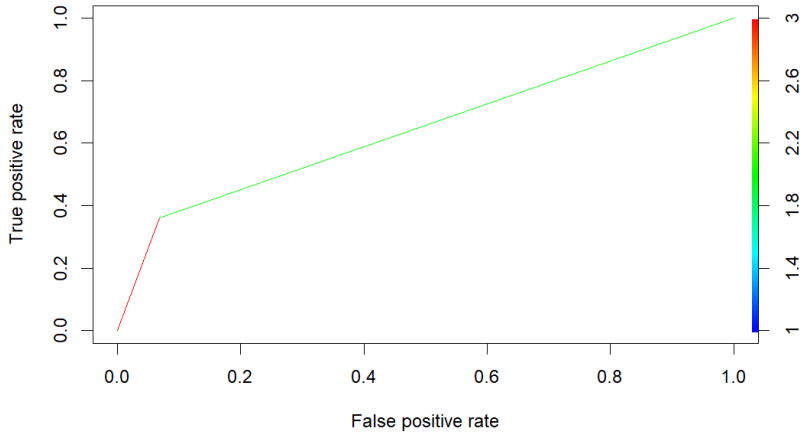


图 10 ROC 曲线

AUC = 0.6461757

**4. 决策树模型**

利用 R 对训练集进行模型构建，得到分类过程如下：

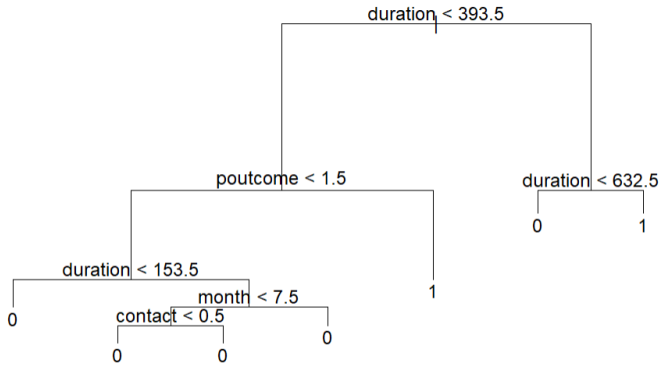


图 11 决策树模型

此模型的根结点为变量 duration，叶结点数量为 7，训练错误率为 11%，平均残差为 16238。

混淆矩阵为：

表 9 混淆矩阵

| | 0 | 1 |
|---|---|---|
| 0 | 5768 | 382 |
| 1 | 327 | 347 |

准确率（Accuracy）= 0.896102

精确率（Precision）= 0.9463495

召回率（Recall）= 0.9378862

灵敏度(Sensitivity) = 0.9378862

特异度(Specificity) = 0.5148368

ROC 曲如下图：

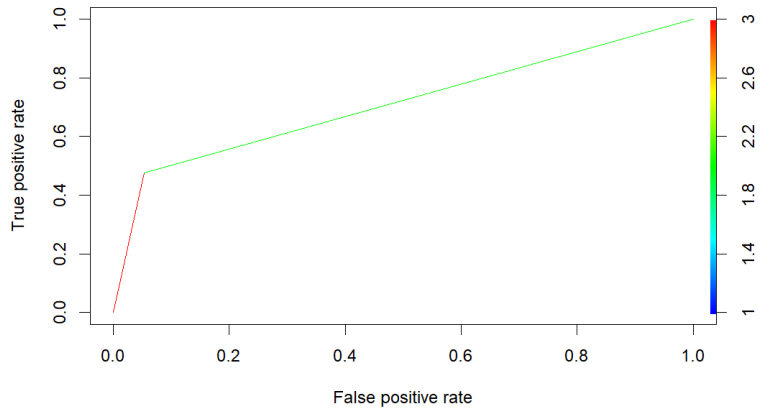


图 12 ROC 曲线

AUC = 0.711172

由于决策树易造成过拟合的问题，下面对分类树进行剪枝处理。

用 R 使用交叉验证 CV 的方法来确定最优的树复杂性，当终端结点数为 4 时，交叉验证错误率最低，共 1817 个交叉验证误差。

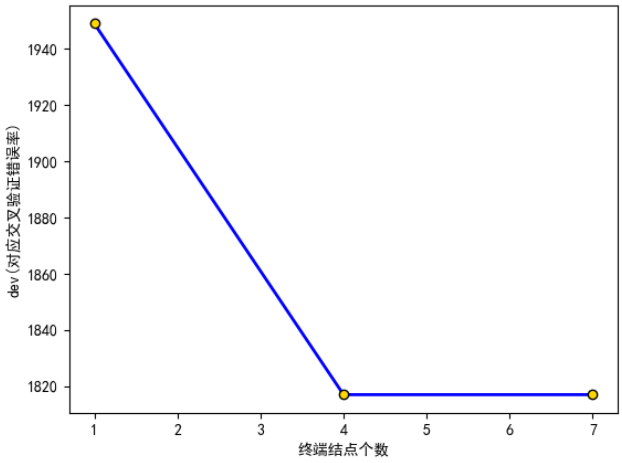


图 13 终端结点与 dev 的关系

下面选择终端结点数为 4，画出剪枝后的分类过程：

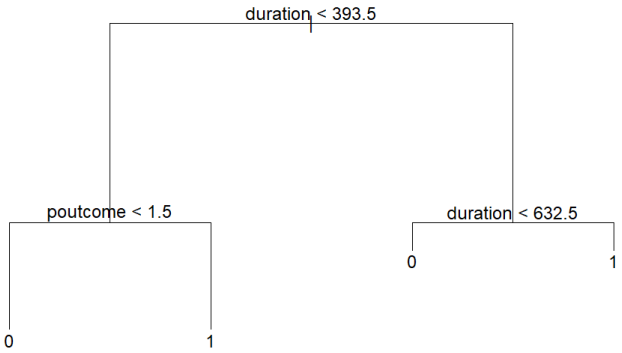


图 14 剪枝后的决策树模型

还是以 duration 为根结点，结合叶结点，可以看出 duration 这一变量在本次分类中的重要性。

混淆矩阵为：

表 10 混淆矩阵

| | 0 | 1 |
|---|---|---|
| 0 | 5768 | 382 |
| 1 | 327 | 347 |

准确率（Accuracy）= 0.896102

精确率（Precision）= 0.9463495

召回率（Recall）= 0.9378862

灵敏度(Sensitivity) = 0.9378862

特异度(Specificity) = 0.5148368

ROC 曲如下图：

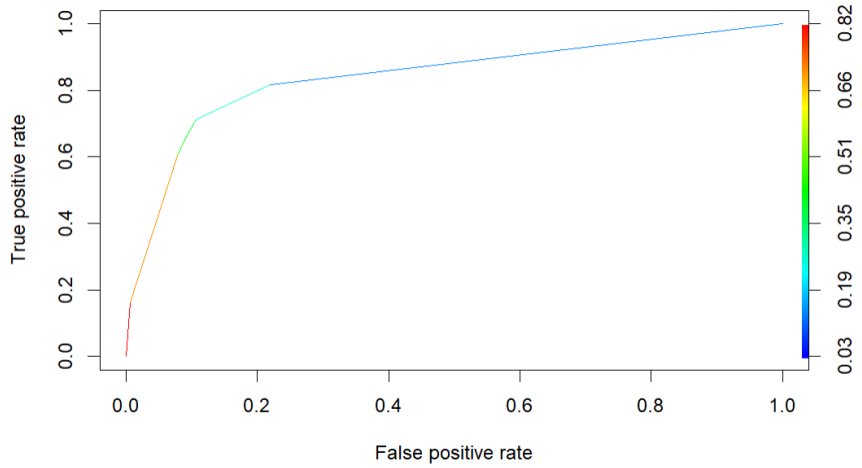


图 15 ROC 曲线

AUC = 0.8425568

综合剪枝前后的结果，可以看出准确率、精确率等指标并无变化，但 AUC 值从 0.711172 上升到了 0.8425568，模型预测的效果有了较好的提升。

## 四、 研究结果与讨论

表 11 四种模型各指标结果比较

| | 模型名称 | | | |
|---|---|---|---|---|
| 指标名称 | Logistic | LDA | KNN | 决策树 |
| 准确率 | 0.8953693 | 0.892585 | 0.8706038 | 0.896102 |
| 精确率 | 0.9717801 | 0.9515997 | 0.9241536 | 0.9463495 |
| 召回率、灵敏度 | 0.916164 | 0.9297852 | 0.9315833 | 0.9378862 |
| 特异度 | 0.5208914 | 0.496587 | 0.3607682 | 0.5148368 |
| AUC | 0.8589598 | 0.8547613 | 0.6461757 | 0.8425568 |

其中，KNN 模型中选取 K=1；决策树模型中采用剪枝后的模型。

### （一） 各指标分析

1. 准确率：

所有模型中，准确率最高的是决策树模型，但与 Logistic 模型和 LDA 模型都相差不大，KNN 模型的准确率相对要低一些。

2. 精确率：

精确率最高的模型是 Logistic 模型，LDA 模型与决策树模型的精确率大小差不多，KNN 显著地低一些。

3. 召回率 /灵敏度：

召回率 /灵敏度最高的是决策树，其次是 KNN、LDA、Logistic 模型。

4. 特异度：

四个模型的特异度都比较低，其中最高的是 Logistic 模型，其次是决策树模型、LDA 模型和 KNN 模型。

5. AUC：

在四个模型中，AUC 最小的是 KNN 模型，仅 0.64，预测效果最差。Logistic 模型和 LDA 模型的 AUC 都在 0.85 左右，决策树的 AUC 稍微低于二者。除 KNN 模型，其他三个模型预测的效果都比较好。

**（二） 综合分析**

在本研究中，银行想要确定那些有更高机会购买长期存款的现有客户，也就是预测得到 y=1 的这一部分客户，银行将把营销工作集中在这些客户身上。对于 y 的预测，我们希望召回率很高，因为银行希望把每一个会购买长期存款的客户都预测出来，并开始下一步的营销工作，针对这些客户，银行有非常大的概率能卖出自己的长期存款产品。以召回率为唯一标准能显著提高真正能购买产品的客户，但是也会增加不会购买的客户被错判的概率。因此，银行在衡量模型时，也要保证模型的精确率，如果精确率过低，银行将对一些不会购买产品的客户进行费时费力的营销，而营销工作本身也有一定的成本，低精确率也会带来一定的损失。对于特异度，较高的特异度意味着银行将不购买产品的客户错判为购买的概率较低，特异度越高，越有利于银行进行下一步的营销工作安排。由于本数据集不平衡，目标变量取值 0-1 所占比例差距较大，约为 89：11，因此准确率在本次分析中不适用于衡量分类效果。 在比较以上几个指标的同时，可以参考 AUC 值，来衡量模型预测的效果。在研究中，可以发现，变量 duration 对模型预测结果的影响较大，这意味着上一次联系客户的时间会对客户的最终决定起到近乎决定性的作用。因此，银行应该扩大联络人员的队伍，并保证对客户联系的频率在一个较高的水平，这样更有利于提高用户的购买率。

对比以上四个模型 ，可以得出以下结论：

Logistic 模型拥有较高的精确率、特异度和 AUC，有利于下一步的营销工作安排。但其

召回率相对其他模型低一些，可能会损失一些本来会购买，但因为缺乏营销，最终并没有购买的客户，造成银行收入损失，但从数值上来看，91.61%的召回率并不算低，此模型是一个好的模型。

LDA 模型拥有较高的精确率和 AUC，召回率和特异度均排在第三位， 但精确率和召回率数值上均较高，也是一个比较好的模型。

KNN 模型的精确率、特异度都是最低，且与其他三个模型有较为明显的差距，其原因应该是受到了维数 p=13 的影响，当维数较大时，KNN 模型表现较差。但 KNN 模型召回率为 93.16%，该算法倾向于预测客户会购买产品，如果银行不在乎营销成本，可以考虑此方法。但一般而言，此模型 AUC 仅有 0.64 左右，不是一个好的预测模型。

决策树模型拥有最高的召回率，和较高的特异度和 AUC，精确率稍低一些，但数值上仍然很高，也是一个不错的模型。

综上，银行决策者可以综合自己的成本、规划考虑，选择 Logistic、LDA 与决策树模型中的一个即可。

# 参考文献

[1]王亮亮.金融危机冲击、融资约束与公司避税[J].南开管理评论,2016,19(01):155-168.

[2]于立勇,詹捷辉.基于 Logistic 回归分析的违约概率预测研究[J].财经研究,2004(09):15-23.DOI:10.16538/j.cnki.jfe.2004.09.002.

[3]张宁,贾自艳,史忠植.使用 KNN 算法的文本分类[J].计算机工程,2005(08):171-172+185.

[4] 统计学习导论——基于 R 应用/(美)詹姆斯 (James，G.) 等著;王星等译. 北京:机械工业出版社，2015.6

[5]栾丽华,吉根林.决策树分类技术研究[J].计算机工程,2004(09):94-96+105.

[6]邹洪侠,秦锋,程泽凯,王晓宇.二类分类器的 ROC 曲线生成算法[J].计算机技术与发展,2009,19(06):109-112.

# 附件

## 一、 变量编码结果

表 1

| Catagory of job |
| --- |
| 0 blue-collar |
| 1 entrepreneur |
| 2 retired |
| 3 admin. |
| 4 student |
| 5 services |
| 6 technician |
| 7 self-employed |
| 8 management |
| 9 unemployed |
| 10 unknown |
| 11 housemaid |

表 2

| Category of marital |
| --- |
| 0 married |
| 1 divorced |
| 2 single |
| 3 unknown |

表 3

| Catagory of education |
| --- |
| 0 basic.9y |
| 1 university.degree |
| 2 basic.4y |
| 3 high.school |
| 4 professional.course |
| 5 unknown |
| 6 basic.6y |
| 7 illiterate |

表 4

| Catagory of default |
| --- |
| 0 unknown |
| 1 no |
| 2 yes |

表 5

| Catagory of housing |
| --- |
| 0 no |
| 1 yes |
| 2 unknown |

表 6

| Catagory of loan |
| --- |
| 0 no |
| 1 yes |
| 2 unknown |

表 7

| Catagory of contact |
| --- |
| 0 cellular |
| 1 telephone |

表 8

| Catagory of month |
| --- |
| 0 nov |
| 1 jul |
| 2 may |
| 3 jun |
| 4 aug |
| 5 mar |
| 6 oct |
| 7 apr |
| 8 sep |
| 9 dec |

表 9

| Catagory of day_of_week |
| --- |
| 0 wed |
| 1 mon |
| 2 tue |
| 3 fri |
| 4 thu |

表 10

| Catagory of poutcome |
| --- |

| 0 nonexistent |
| --- |
| 1 failure |
| 2 success |

表 11

| Catagory of y |
| --- |
| 0 no |
| 1 yes |

## 二、 代码展示

**1. 数据预处理：Python 代码**

```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import os
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
train_data = pd.read_csv('D://大三上//统计学习//期末报告//数据//银行数据//new_train.csv')
df = pd.DataFrame(train_data)
df.head(5)
%matplotlib qt5
# percentage of class present in target variable(y)
print("percentage of NO and YES\n",df["y"].value_counts()/len(df)*100)
#查看缺失
data.isnull().sum()
#去除变量
data.drop(columns=["pdays", "previous"], axis=1, inplace=True)
#处理离群值
data.describe()
# compute interquantile range to calculate the boundaries
lower_boundries= []
upper_boundries= []
for i in ["age", "duration", "campaign"]:
    IQR= data[i].quantile(0.75) - data[i].quantile(0.25)
    lower_bound= data[i].quantile(0.25) - (1.5*IQR)
    upper_bound= data[i].quantile(0.75) + (1.5*IQR)
    print(i, ":", lower_bound, ",",  upper_bound)
    lower_boundries.append(lower_bound)
```

```
        upper_boundries.append(upper_bound)
# replace the all the outliers which is greater then upper boundary by upper boundary
j = 0
for i in ["age", "duration", "campaign"]:
        data.loc[data[i] > upper_boundries[j], i] = int(upper_boundries[j])
        j = j + 1
#编码
dftObjcat = dftObject.columns
for i in dftObjcat:
        print(f'Catagory of {i}')
        catlist = dftObject[i].unique()
        for j, val in enumerate(catlist):
                dftobjfinal = dftObject[i].replace({val:j+1},inplace=True)
                #print(dftobjfinal)
                print(j,val)
```

**2. 相关系数热力图：Python 代码**

```
 import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns   # 可视化 w
inputdata = pd.read_excel("D://大三上//统计学习//期末报告//数据//银行数据//trainnoy.xls")
plt.rcParams['font.sans-serif'] = ['SimHei']   # 用来正常显示中文标签
plt.rcParams['axes.unicode_minus'] = False   # 用来正常显示负号
df = inputdata.copy()
_, ax = plt.subplots(figsize=(12, 10))   # 分辨率 1200×1000
corr = df.corr(method='spearman') # 斯皮尔曼秩相关系数
cmap = sns.diverging_palette(220, 10, as_cmap=True)
_ = sns.heatmap(
        corr,   # 使用 Pandas DataFrame 数据，索引/列信息用于标记列和行
        cmap=cmap,   # 数据值到颜色空间的映射
        square=True,   # 每个单元格都是正方形
        cbar_kws={'shrink': .9},   # `fig.colorbar`的关键字参数
        ax=ax,   # 绘制图的轴
        annot=True,   # 在单元格中标注数据值
        annot_kws={'fontsize': 12})   # 热图，将矩形数据绘制为颜色编码矩阵
plt.title("各解释变量之间的相关性强弱", fontsize=20)
plt.show()
```

**3. Logistic 回归：R 代码**

```
obj=glm(y~age+job+marital+education+default+housing+loan+contact+month+day_of_week+
duration+campaign+poutcome,data = train)
summary(obj)
```

```
fit.reduced <- glm(y ~ age + job + marital + education + contact+month+duration+campaign+poutcome ,family=binomial(),data=train)
summary(fit.reduced)
coef(obj)
mydata = train
mydata$y[mydata$y == '1'] = 1
mydata$y[mydata$y == '0'] = 0
mydata$y = as.numeric(mydata$y)
index = sample(x = 1:2,size = nrow(mydata), replace = TRUE, prob = c(0.7,0.3))
train = mydata[index == 1, ]
test = mydata[index == 2, ]
logistic.model = glm(y~., data = train, family = binomial(link = 'logit'))
train_predict0 = predict(logistic.model, train, type='response')
train_predict = ifelse(train_predict0>0.5, 1, 0)
test_predict0 = predict(logistic.model, test, type='response')
test_predict = ifelse(test_predict0>0.5, 1, 0)
predict_value=test_predict0
true_value=test[,14]
```

**4. LDA：R 代码**

```
library(MASS)
lda.fit=lda(y~age+job+marital+education+default+housing+loan+contact+month+day_of_week+duration+campaign+poutcome,data=train)
lda.pred=predict(lda.fit,test)
names(lda.pred)
lda.pred
lda.class=lda.pred$class
#plot(lda.fit)
#lda.fit
lda.class#预测
test$y#实际
t=table(lda.class,test$y)
```

**5. KNN：R 代码**

```
library(class)
set.seed(1)  #设置一个随机种子
knn.pred = knn(train, test,train$y, k = i)
t=table(test$y,knn.pred)
correct = rep(0,15)
for(i in 1:15){
  fit_pre = knn(train, test,train$y,k=i)
  correct[i] = mean(fit_pre ==  test$y)
```

```
}
```

**6. 决策树：R 代码**

```
 #install.packages("tree")
library(tree)
library(rpart)
#install.packages("rpart.plot")
library(rpart.plot)
#install.packages("rattle")
library(rattle)
chose=ifelse(train$y>0.5,"1", "0")
bank=data.frame(train,chose)
tree.bank=tree(chose~.-y,bank)
summary(tree.bank)
plot(tree.bank)
text(tree.bank,pretty=0)
tree.bank=tree(chose~.-y,bank)
tree.pred=predict(tree.bank,test,type='class')
t=table(tree.pred,test$y)
set.seed(3)#剪枝
cv.bank=cv.tree(tree.bank,FUN = prune.misclass)
names(cv.bank)
cv.bank
prune.bank=prune.misclass(tree.bank,best=4)
plot(prune.bank)
text(prune.bank,pretty=0)
tree.pred=predict(prune.bank,test,type="class")
t=table(tree.pred,test$y)
```

**7. 计算各指标+可视化 ROC：R 代码**

```
#混淆矩阵，显示结果依次为 TP、FN、FP、TN
t=table(test_predict,test$y)
t
tp <- t[1, 1]
tn <- t[2, 2]
fp <- t[2, 1]
fn <- t[1, 2]
print(accuracy <- (tp + tn)/(tp + tn + fp + fn))
print(precision <- tp/(tp + fp))
print(recall <- tp/(tp + fn))
print(sensitivity <- tp/(tp + fn))
print(specificity <- tn/(tn + fp))
```

```
#install.packages("caret", dependencies = c("Depends", "Suggests"))
library(pROC)
library(ggplot2)
library(magrittr)
#install.packages("ROCR")
library(ROCR)
pred <- prediction(predict_value,true_value)     #预测值(0.5 二分类之前的预测值)和真实值
performance(pred,'auc')@y.values                 #AUC 值
perf <- performance(pred,'tpr','fpr')    #y 轴为 tpr(true positive rate),x 轴为 fpr(false positive rate)
plot(perf,colorize=TRUE)
```

**8. 正态性检验：spss 代码**

```
NEW FILE.
DATASET NAME  数据集1 WINDOW=FRONT.
GET DATA
   /TYPE=XLS
   /FILE='C:\Users\tlh\Desktop\train.xls'
   /SHEET=name 'Sheet1'
   /CELLRANGE=FULL
   /READNAMES=ON
   /DATATYPEMIN PERCENTAGE=95.0.
EXECUTE.
DATASET NAME  数据集2 WINDOW=FRONT.
DATASET CLOSE  数据集1.
EXAMINE VARIABLES=age job marital education default housing loan contact month
day_of_week duration
      campaign poutcome
   /PLOT BOXPLOT STEMLEAF NPPLOT
   /COMPARE GROUPS
   /STATISTICS DESCRIPTIVES
   /CINTERVAL 95
   /MISSING LISTWISE
   /NOTOTAL.
```

## 三、 数据展示

表 12

| age | job | marital | education | default | housing | loan | contact | month | day_of_we | duration | campaign | pdays | previous | poutcome | y |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 49 | blue-collar | married | basic.9y | unknown | no | no | cellular | nov | wed | 227 | 4 | 999 | 0 | nonexistent | no |
| 37 | entrepreneu | married | iversity.degı | no | no | no | telephone | nov | wed | 202 | 2 | 999 | 1 | failure | no |
| 78 | retired | married | basic.4y | no | no | no | cellular | jul | mon | 1148 | 1 | 999 | 0 | nonexistent | yes |
| 36 | admin. | married | iversity.degı | no | yes | no | telephone | may | mon | 120 | 2 | 999 | 0 | nonexistent | no |
| 59 | retired | divorced | iversity.degı | no | no | no | cellular | jun | tue | 368 | 2 | 999 | 0 | nonexistent | no |
| 29 | admin. | single | iversity.degı | no | no | no | cellular | aug | wed | 256 | 2 | 999 | 0 | nonexistent | no |
| 26 | student | single | basic.9y | no | no | no | telephone | aug | wed | 449 | 1 | 999 | 0 | nonexistent | yes |
| 30 | blue-collar | married | basic.4y | no | yes | no | cellular | nov | wed | 126 | 2 | 999 | 0 | nonexistent | no |
| 50 | blue-collar | married | basic.4y | unknown | no | no | telephone | may | fri | 574 | 1 | 999 | 0 | nonexistent | no |
| 33 | admin. | single | high.school | no | yes | no | cellular | jul | tue | 498 | 5 | 999 | 0 | nonexistent | no |
| 44 | services | divorced | high.school | no | yes | no | cellular | jul | mon | 158 | 5 | 999 | 0 | nonexistent | no |
| 32 | technician | married | iversity.degı | no | yes | no | telephone | may | fri | 93 | 5 | 999 | 0 | nonexistent | no |
| 26 | elf-employe | single | essional.coı | no | yes | no | cellular | jul | thu | 71 | 1 | 999 | 0 | nonexistent | no |
| 43 | managemen | married | iversity.degı | no | no | yes | telephone | jul | thu | 203 | 1 | 999 | 0 | nonexistent | no |
| 56 | blue-collar | married | basic.9y | no | no | no | cellular | may | thu | 369 | 1 | 999 | 0 | nonexistent | no |
| 40 | blue-collar | married | basic.9y | no | yes | no | cellular | may | wed | 954 | 1 | 999 | 0 | nonexistent | yes |
| 32 | admin. | divorced | iversity.degı | no | yes | no | cellular | aug | tue | 105 | 1 | 999 | 0 | nonexistent | no |
| 47 | technician | single | essional.coı | unknown | no | no | telephone | may | tue | 148 | 5 | 999 | 0 | nonexistent | no |
| 50 | services | single | basic.4y | no | yes | no | telephone | may | tue | 98 | 9 | 999 | 0 | nonexistent | no |
| 34 | admin. | single | iversity.degı | no | no | yes | cellular | mar | tue | 288 | 2 | 3 | 1 | success | yes |
| 46 | services | married | unknown | no | no | no | cellular | aug | tue | 177 | 1 | 999 | 0 | nonexistent | no |
| 39 | blue-collar | married | basic.4y | no | no | no | cellular | may | thu | 155 | 1 | 999 | 0 | nonexistent | no |
| 41 | admin. | divorced | basic.6y | no | no | no | telephone | jul | mon | 141 | 1 | 999 | 0 | nonexistent | no |
| 30 | technician | single | unknown | no | yes | no | telephone | may | mon | 144 | 3 | 999 | 0 | nonexistent | no |
| 55 | unemployed | divorced | iversity.degı | no | no | no | cellular | mar | tue | 212 | 3 | 6 | 3 | success | yes |
| 33 | blue-collar | single | basic.4y | no | yes | no | cellular | oct | fri | 146 | 1 | 999 | 1 | failure | no |
| 46 | services | married | high.school | no | no | no | cellular | apr | wed | 325 | 2 | 999 | 0 | nonexistent | no |
| 38 | blue-collar | divorced | high.school | no | yes | yes | cellular | nov | wed | 291 | 1 | 999 | 0 | nonexistent | no |
| 36 | admin. | divorced | iversity.degı | unknown | yes | yes | cellular | jul | wed | 103 | 1 | 999 | 0 | nonexistent | no |

注：本数据集共有 32952 行，此处仅展示前 36 行数据。