郑州大学毕业论文

题目：基于生存分析和机器学习对手机客户流失问题的研究

指导教师：李　锋　　职称：副教授

学生姓名：郭林怡　学号：20142110209

专　　业：数学与应用数学

院（系）：数学与统计学院

完成时间：2018 年 5 月 27 日

2018 年 5 月 27 日

摘要

面对日趋饱和的电信市场与激烈的行业竞争，如何防止现有的客户流失成为各电信运营商关心的问题，因此，对手机客户流失因素的分析以及预测模型的建立具有重要的实际意义。文章首先交代了手机客户流失问题的研究背景以及目前国内外关于该问题的研究情况，然后对与有关的概念和采用的研究方法进行了简单的介绍。文章的主体部分采用了生存分析和机器学习两方面的知识分别解决针对该问题的分析与建模。Cox 比例风险回归模型给出了不同因素对客户流失的影响分析，其中，客户是否购买额外的电信服务产品以及是否签订合约对其流失有重要影响。Logistic 回归和决策树提供了预测客户是否流失的模型，交叉验证的结果表明，各模型预测效果都还不错，相对而言，C5.0 决策树表现更好一些。文章的最后给出了关于该问题未来研究的一些方向。

关键词：用户流失；生存分析；机器学习；Logistic 回归；决策树

Abstract

Facing gradually saturated telecommunication market and furious industrial competition, telecom operators begin to pay more attention to how to prevent their current users from losing, so it is important to analyze related factors and build models for prediction. This paper illustrated the backgrounds of the loss problem of mobile phone users and the current research situation at home and abroad firstly, and then introduced some concepts and research methods applying in the paper briefly. The main part adopted survival analysis and machine learning to solve factor analysis and model construction respectively. Cox proportional hazards model outputs an analysis of the influence of different factors towards users' loss. Among several factors, whether users bought additional telecom products and whether they signed a contract play significant roles. Logistic regression and decision tree provide models for loss prediction, and the results of cross validation showed that every model does make sense. Comparatively, C5.0 decision tree have a better performance. Finally, the paper gives some possible research orientations about this problem in the future.

Key words: Customer Loss; Survival Analysis; Machine Learning; Logistic Regression; Decision Tree

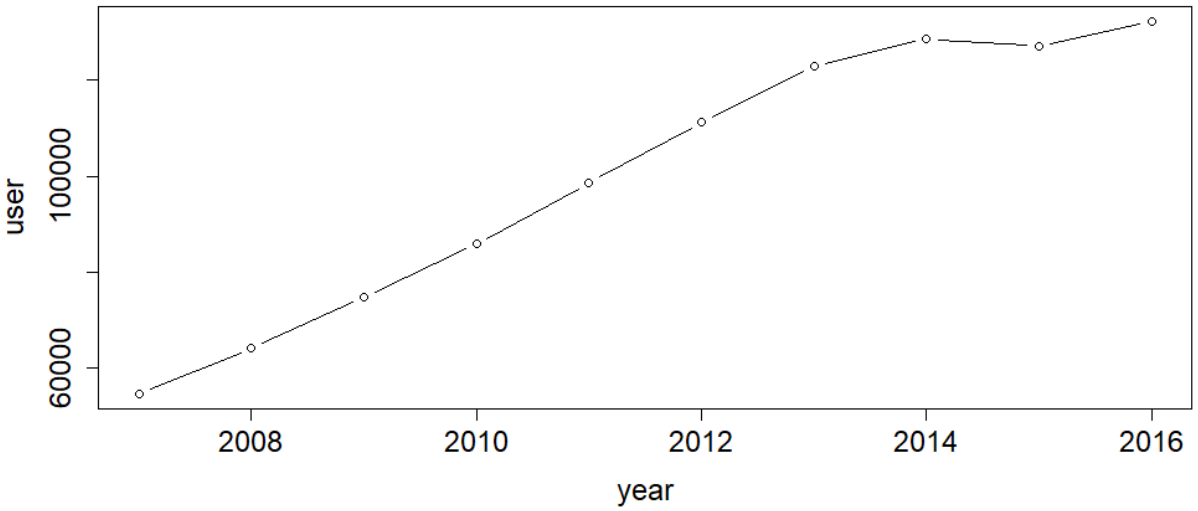
目录

1 引言

1.1 问题背景

自 1987 年我国第一台手机面世后，我国的手机用户就不断增多，据国家统计局数据显示，截止 2016 年年底，我国共有高达 132193.4 万移动电话用户。但是，随着手机的日益普及，手机号注册量越来越少，表 1.1 给出了 2007-2016 年的移动电话用户总量变化（数据来自国家统计局）。

图 1.1 近 6 年手机用户数量变化情况



不难发现，2013 年以来，手机用户的增长态势已经放缓，也就是说，三大电信企业（移动，联通和电信）如果想要保持业绩或实现业务增长，发展新用户的作用就显得很有限。

电信行业竞争日益激烈，面对日趋饱和的电信市场，为了争夺市场份额，各大运营商不断推出新的套餐和各种优惠政策来吸引潜在的新客户。我国的手机客户数量巨大，其中中低端客户占了很大比例，而且这部分客户在各大运营商之前的切换又较为频繁。

证据表明，在激烈的竞争市场下，发展新客户的成本往往高于维护老客户的成本[1]。著名的“二八”定律表明，企业的 20%的高价值用户将提供了企业总盈利的 80%左右[2]。因此，如何维系已经拥有的老客户，特别是高价值客户，就显得格外重要。

1.2 研究目的及意义

以某电信公司的 4975 条数据记录为样本，用生存分析的方法分析一系列因素（比如用户套餐金额和额外通话时长）对用户使用时长以及用户是否流失的影响，希望通过机器学习的知识建立合适的模型来预测单个用户在未来是否会流失，与此同时，借助分析所得到的结论，指导电信企业制定合适的策略来避免客户流失。

直观上，这个研究可以保证甚至提高电信企业的利润。长远来看，维系好老客户有助于提高企业的整体形象，并方便后期新业务的推广，从而进一步提高企业利润。

1.3 国内外研究情况

对客户流失问题的研究始于国外，比如 Reicheld 和 Sasser 于 1999 年就指出客户流失和忠诚度有很大关系[15]；Drew 等人于 2001 年发表文章称人工神经网络对客户流失问题有更好的预测[16]；Louis 在 2002 年发表文章，对比了 Logistic 回归和决策树两种模型各自的预测效果[17]。虽然我国移动通讯产业起步较晚，但由于受到巨大的市场需求的刺激，关于客户流失问题的各项研究发展很快[18]，从最初柳兰屏和曾煌（2003）对客户细分的研究[19]，到林向阳（2010）通过数据挖掘对用户流失进行分析[2]，相关领域的研究不断取得新的突破。

综合来看，国内外大多数移动企业都是围绕 CRM（Customer Relationship Management，客户关系管理）来构建客户流失管理的数据库及相关的预测模型[3]，比如早期就比较成功的 Verizon 公司和英国电信集团（BT）。但是由于科技的日新月异，无论国内还是国外，该领域的研究还在持续进行中。

2 手机客户流失问题的分析方法简介

2.1 客户流失的定义

客户流失(Customer Loss)是指客户因某种原因和企业终止合作的现象。在电信行业中，客户流失一般指客户因某种原因和电信运营企业解除服务合同，即客户停止或减少消费目前正在使用的电信产品或服务，选择其他电信企业的产品或服务，或者该电信运营企业其他的代替性电信产品或服务，或者终止使用任何一项电信服务[4]。

客户流失主要表现为四个方面：

(1)更换所使用的电信运营企业，即退网；

(2)高消费套餐转低消费套餐；

(3)ARPU 值（平均月消费）降低；

(4)网内弃卡换号，即在不更换运营商的前提下进行换号。

2.2 客户关系管理

引言中，我们提到目前大多数电信运营企业都是根据 CRM 来构建数据库以及预测模型。事实上，在 1980 年初便有所谓的“接触管理”（Contact Management）专门收集客户与公司联系的所有信息，到 1990 年则演变成了包括电话服务中心支持资料分析的“客户关怀”（Customer Care）。我们可以把 CRM 定义[6]为：企业为提高核心竞争力，利用相应的信息技术以及互联网技术协调企业与顾客间在销售、营销和服务上的交互，从而提升其管理方式，向客户提供创新式的个性化的客户交互和服务的过程[5]。其最终目标是吸引新客户、保留老客户以及将已有客户转为忠实客户，增加市场。

2.3 生存分析简介

生存分析（Survival Analysis）是将事件的结果和出现此结果所经历的时间结合起来分析的现代统计方法，也称之为风险模型或持续模型(Hazard Model/Duration Model)[7]。生存分析方法最初是从医学寿命资料的统计分析中发展起来的，医学研究中，为了了解某种疾病的预测、评价治疗方法的优劣或观察预防保健措施的效果等，常需对研究对象进行追踪观察，以获得必要的数据[8]。由于生存分析源于对研究对象生存时间或寿命的研究，因此也被称为生存时间分析。生存分析中提到的生存时间是广义上的，它可以是通常意义上的时间，比如癌症病人确诊后的存活时间，也可以是汽车总行程的公里数，工厂机床齿轮转动的圈数等。

处理生存分析问题一般包含下列三个过程：

(1)描述生存过程，比如生存时间分布特点，生存率等，常用 Kaplan-Meier 法（简称 K-M 法）和寿命表法；

(2)比较生存过程，通过比较生存率和标准误差等因素来判断不同组间生存过程是否存在差异，常用 log-rank 检验；

(3)影响生存时间因素分析，通过建立模型判断保护因素和不利因素，以及因素作用大小，常用 Cox 比例风险回归模型（Cox Proportional Hazards Model）。

2.4 机器学习简介

机器学习（Machine Learning），顾名思义，是专门研究如何使用计算机来模拟或实现人类活动的一门学科，其最早的应用大概是 1959 年美国的 Samuel 所发明的西洋棋程序。我们希望计算机可以在已有经验的帮助下，通过各种程序来提升能力，这样在下一次执行相同或相似的问题时,就可以有更高的效率。目前,ML 还没有一个统一且准确的定义。在各类定义中，由卡耐基梅隆大学（Carnegie Mellon University）的 Tom Mitchell 在其著作《Machine Learning》给出的定义被广泛接受：一个计算机程序从经验 E 中学习任务 T 和性能衡量标准 P 的相关方面，如果任务 T 的性能（由 P 进行度量）能够由 E 得到改善，则称该方法为机器学习[9]。

由于分类方式的不同，机器学习有很多种分类体系，其中较为常见的当属基于学习方式的分类：

(1)监督式学习(Supervised Learning),常见算法有逻辑回归(Logistic Regression)和反向传递神经网络（Back Propagation Neural Network）；

(2)无监督学习（Unsupervised Learning），常见算法有 k-Means 聚类算法；

(3)强化学习（Reinforcement Learning），常见算法有时间差学习（Temporal Difference Learning）[10]。

另外，经常用到的方法还有决策树，支持向量机（Support Vector Machine, SVM）以及贝叶斯分类方法等。从收集的国内相关论文的情况来看，客户流失分析较为常用的有四种方法,分别是决策树方法,神经元网络方法,Logistic 回归方法和贝叶斯分类方法[2]。

3 数据准备

3.1 数据来源

本文所分析的数据来自微信公众号狗熊会，数据中的 4975 条真实观测均来自国内某电信运营商，每条观测值来自一个手机号码的某个年度。

3.2 变量引入

本文原始数据中自变量有 7 个，详细信息见下表。

表 3.1 自变量信息

变量名称	变量类型	变量名
套餐金额	哑元变量	FEE
额外通话时长	连续变量	EXTIME
额外流量	连续变量	EXDATA
改变行为	哑元变量	CHANGE
服务合约	哑元变量	CONTRACT
关联购买	哑元变量	PRODUCTS
集团用户	哑元变量	GROUP

注：(1)套餐金额：1,2,3 分别代表套餐金额在 0-96 元，96-255 元以及 255 元以上；

(2)额外通话时长是指用户每月使用实际通话时长超出套餐所含通话时长的部分，其大小是观测期间每月额外通话时长的平均值（单位：分钟），额外流量与其类似（单位：兆）；

(3)改变行为：用户在使用期间是否更换套餐，1=是，0=否；

(4)服务合约：用户是否和电信公司签订合约，1=是，0=否；

(5)关联购买：用户在该电信公司办理其他电信服务的数量，主要是固话和宽带服务，取值有 0,1,2 三种情况；

(6)集团用户：在某一集团内，办理该业务的用户互相拨打有一定优惠，1=是，0=否。

响应变量（因变量）为用户使用月数以及用户是否流失，两者类型均为哑元变量，前者为取值在[1,25]内的整数，后者为 0-1 二元变量（1 代表用户流失）。

3.3 数据预处理

在对数据进行正式建模之前，我们先对其进行一些基本的统计分析。

在 4975 条观测数据中，月套餐金额（单位：元）在 1-96，96-225 和 225 以上的人数比为 4719 ： 225 ： 31，显而易见，普通用户占据了该电信公司所有用户的很大比例。

考虑到之前我们提到：用户是否在在使用期间更换套餐，是否签订服务合约，是否为集团用户，同时办理的业务数量，以及我们最关心的客户的流失情况，通过表 3.2 可以直接反映出来：

表 3.2 各变量数量分布情况

	0	1	2
改变行为	4869	106	NULL
服务合约	3755	1220	NULLL
关联购买	4819	76	80
集团用户	3844	1131	NULL
流失用户	1081	3894	NULL

注：NULL 表示该变量在对应数值下没有定义。

对于额外通话时长，额外流量以及使用月数，我们用其直方图表示更为直观：

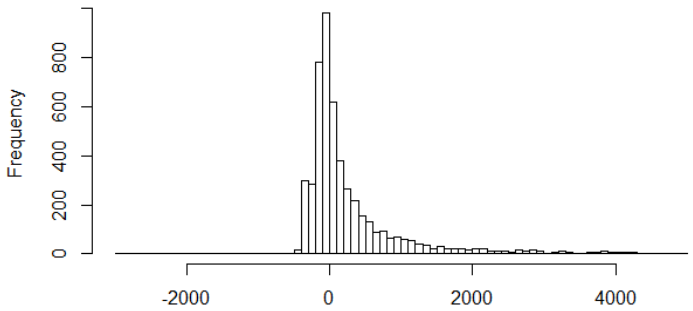


图 3.1 用户额外通话时长分布直方图

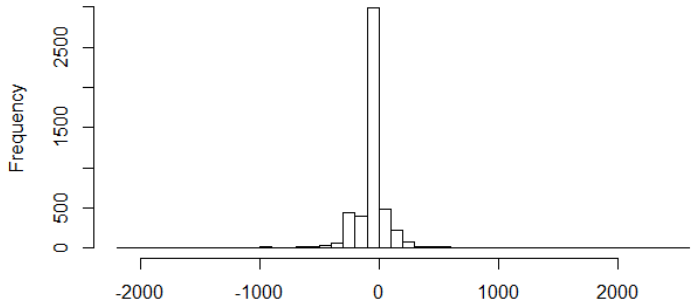


图 3.2 用户额外流量分布直方图

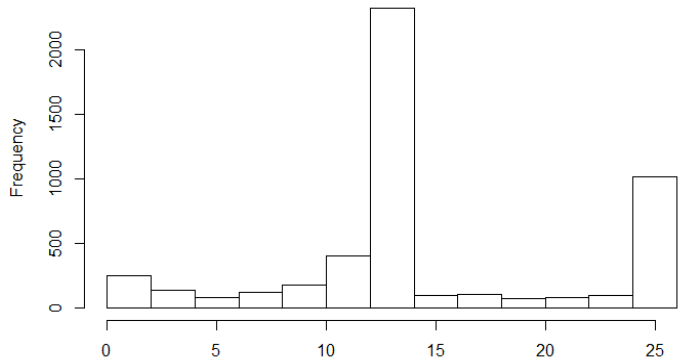


图 3.3 用户使用月数分布直方图

用 summary()函数对其分析后，得到下表

表 3.3 summary 输出结果

	Min.	1ST Qu.	Median	Mean	3rd Qu.	Max.
EXTIME	-2828.3	-126.7	13.5	258.5	338.7	4314.0
EXDATA	-2189.9	-74.3	-59.7	-71.6	-25.8	2568.7
DURATION	1.0	13.0	13.0	14.77	19.0	25.0

另外，根据其分布的特点，我们还计算了额外通话时长和额外流量的偏度，分别为 2.698157 和-0.4633497。

通过上述数据，我们可以有以下两点结论：

(1)该电信公司目前用户流失情况严重，且很大一部分用户流失发生在使用 12-14 个月期间；

(2)额外通话时间呈现明显的偏态分布，额外流量也呈现出偏态分布的特点，且相当一部分用户存在通话时长和流量“用不完”的情况。

显然，结论（1）符合我们正常的认知，比如一些购买合约机的用户在合约到期后就会更换其他的手机号。另外通过结论（2），我们可以猜测该电信公司之所以用户流失现象严重可能是客户由于通话时长和流量用不完而更换了其他电信公司更为廉价的产品。

由于我们的目的是为了确定这些因素和用户流失的关系，所以为了更好的理解数据，我们将 4975 条观测值在上述分析的基础之上，再细分为流失和没有流失两类，便可以得到表 3.4：

表 3.4 五个哑元变量的细分情况

	0		1		2	
流失情况	0	1	0	1	0	1
套餐金额	920	3799	135	90	26	5
改变行为	1024	3845	57	49	NULL	
服务合约	415	3340	666	554	NULL	
关联购买	960	3859	61	15	60	20
集团用户	459	3385	622	509	NULL	

注：此处我们将之前定义的套餐 1,2,3 分别赋给 0,1,2。

为了更直观的显示，我们选取两个影响比较明显的因素（CONTRACT 和 GROUP）进行绘图:

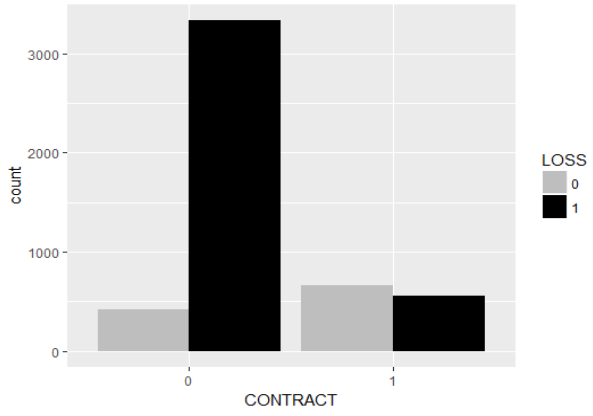


图 3.5 是否签订服务合约的细分情况

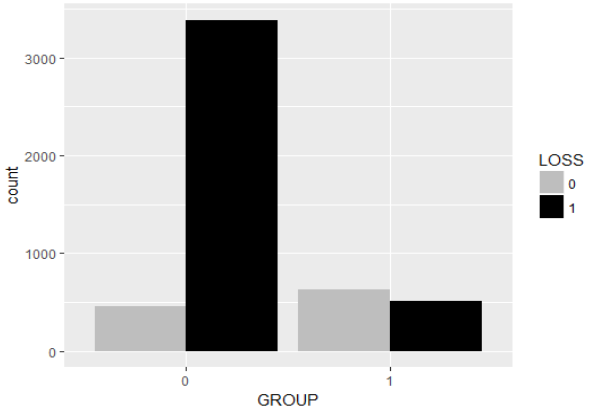


图 3.6 是否为集团用户的细分情况

下面是两个连续自变量细分情况的分布直方图：

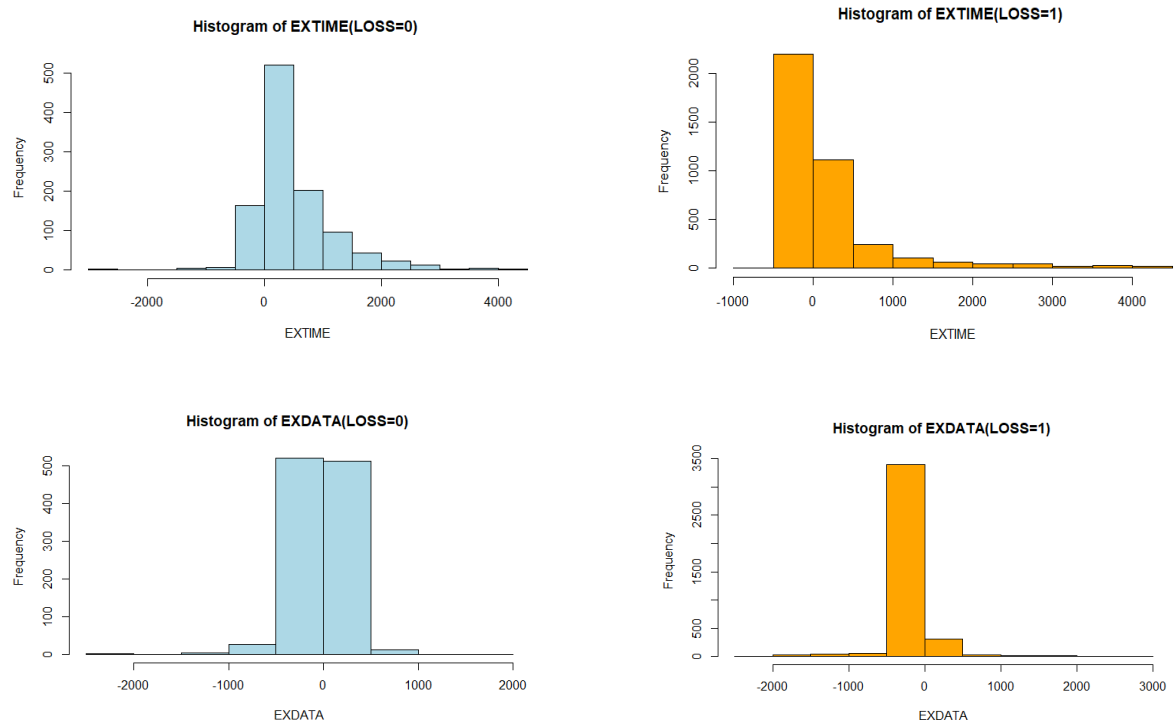


图 3.7 额外通话时长和额外流量细分后的分布情况

从上述图表中，我们不难发现当用户满足表 3.4 中的某一个或多个变量大于 0 时，这类用户流失的比例就会小很多，即：当我们办理的手机业务越多或消费的套餐越高，我们越不会轻易更换运营商，从客户忠诚度角度来说，是作为消费者的我们对该公司或某类产品形成了“依附性”偏好[11]。另外，对于通话时长和使用流量，许多流失用户都存在“过剩”的情况，尤其是在流量的使用方面，“过剩”的情况更加明显。这一现象也印证了之前我们根据结论（2）得出的猜测。

对数据的基本操作到这里就告一段落，本小节我们主要是借助图表从直观上初步理解或猜想变量间的相互关系。

在接下来的分析中，我们会从两个角度来考虑如何防止客户流失：首先，我们用生存分析的方法来研究造成用户流失的因素，根据 Cox 回归的输出结果，电信公司可以制定合适的营销战略；其次，我们用机器学习中常见的 Logistic 回归和决策树进行建模，然后电信公司可以在模型的帮助下，根据用户的个人信息来判断其是否会流失并作出相应决断。

为了方便研究，在接下来的内容中，我们将套餐金额和关联购买两项设置成虚拟变量，分别令 0=套餐 1，1=套餐 2+套餐 3；0=关联购买 0，1=关联购买 1+关联购买 2。

4 基于生存分析的研究

4.1 生存过程

1958 年，Kaplan 和 Meier 提出了关于生存函数的乘积极限法，用公式可以表示为：

$$S(t) = P\{T \geq t\} = \prod_{X_{(i)} \leq t} p_i = \prod_{X_{(i)} \leq t} S_{(t|t_{i-1})}$$

其中，$X_{(i)}$是生存时间$\{X_i\}$的顺序统计量（$i = 1 \ldots m$，m是样本个数），p_i可以看做每个阶段（可以理解成死亡点）处的生存率。

在 R 中调用“survival”包，利用 survfit（）语句我们可以得到下列图表：

表 4.1 每个阶段的生存情况

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
1	4975	164	0.967	0.00253	0.962	0.972
2	4811	90	0.949	0.00312	0.943	0.955
3	4721	80	0.933	0.00355	0.926	0.940
4	4641	56	0.922	0.00381	0.914	0.929
5	4585	43	0.913	0.00400	0.905	0.921
6	4542	40	0.905	0.00416	0.897	0.913
7	4502	51	0.895	0.00435	0.886	0.903
8	4450	71	0.880	0.00460	0.871	0.889
9	4379	87	0.863	0.00488	0.853	0.873
10	4292	88	0.845	0.00513	0.835	0.855
11	4203	112	0.823	0.00542	0.812	0.833
12	4091	291	0.764	0.00602	0.752	0.776
13	3799	1973	0.367	0.00684	0.354	0.381
14	1826	353	0.296	0.00648	0.284	0.309
15	1472	8	0.295	0.00647	0.282	0.308
16	1464	87	0.277	0.00635	0.265	0.290
17	1377	57	0.266	0.00626	0.254	0.278
18	1319	44	0.257	0.00620	0.245	0.269
19	1271	48	0.247	0.00612	0.235	0.259
20	1223	25	0.242	0.00608	0.230	0.254
21	1197	40	0.234	0.00601	0.223	0.246
22	1154	28	0.228	0.00596	0.217	0.240
23	1116	31	0.222	0.00590	0.211	0.234
24	1072	27	0.216	0.00585	0.205	0.228

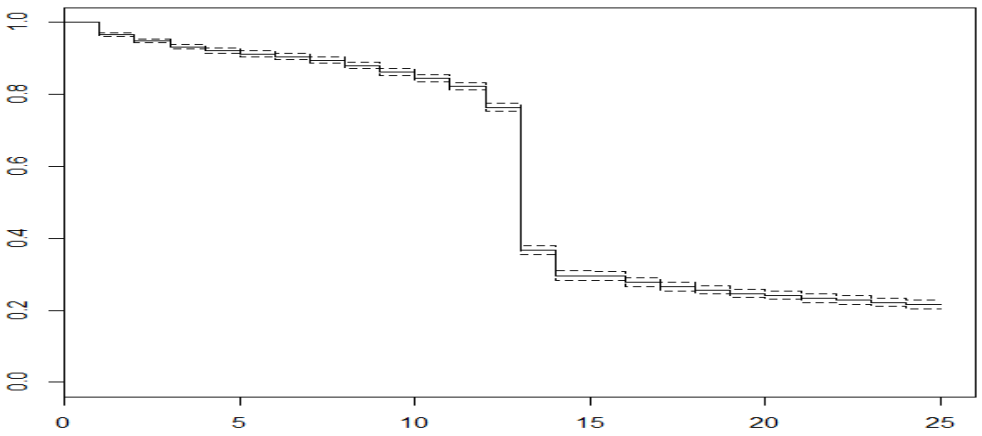


图 4.1 4975 条数据的 k-m 曲线

表中第一列代表使用时长，第二列是我们计算p_i将用到的一个数值，第三列代表该使用时长对应的样本的频数，第四列为该使用时长对应的生存率的估计。此处的分析再次证明了我们之前数据预处理中得到的结论：该电信企业的客户流失主要发生在 13 月左右，且最终客户的生存率在 0.2 左右。

4.2 生存过程的比较

这里，我们只简单选择 EXPRODUCTS（是否办理额外的电信业务）进行简单的分析：

表 4.2 是否办理额外电信服务的用户群体各自的生存过程

EXPRODUCTS=0

time	n.risk	n.event	survival	std.err
1	4819	162	0.966	0.00260
2	4657	90	0.948	0.00321
3	4567	77	0.932	0.00363
4	4490	56	0.920	0.00391
5	4434	43	0.911	0.00410
6	4391	40	0.903	0.00427
7	4351	51	0.892	0.00447
8	4299	68	0.878	0.00471
9	4231	85	0.861	0.00499
10	4146	88	0.842	0.00525
11	4057	111	0.819	0.00554
12	3946	290	0.759	0.00616
13	3655	1971	0.350	0.00687
14	1684	352	0.277	0.00645
15	1331	8	0.275	0.00643
16	1323	86	0.257	0.00630
17	1237	57	0.245	0.00620
18	1179	40	0.237	0.00613
19	1135	48	0.227	0.00604
20	1087	23	0.222	0.00599
21	1063	36	0.215	0.00592
22	1024	25	0.209	0.00587
23	990	29	0.203	0.00580
24	949	23	0.198	0.00575

EXPRODUCTS=1

time	n.risk	n.event	survival	std.err
1	156	2	0.987	0.00901
3	154	3	0.968	0.01410
8	151	3	0.949	0.01766
9	148	2	0.936	0.01961
11	146	1	0.929	0.02050
12	145	1	0.923	0.02133
13	144	2	0.910	0.02288
14	142	1	0.904	0.02360
16	141	1	0.897	0.02429
18	140	4	0.872	0.02677
20	136	2	0.859	0.02787
21	134	4	0.833	0.02984
22	130	3	0.814	0.03115
23	126	2	0.801	0.03196
24	123	4	0.775	0.03347

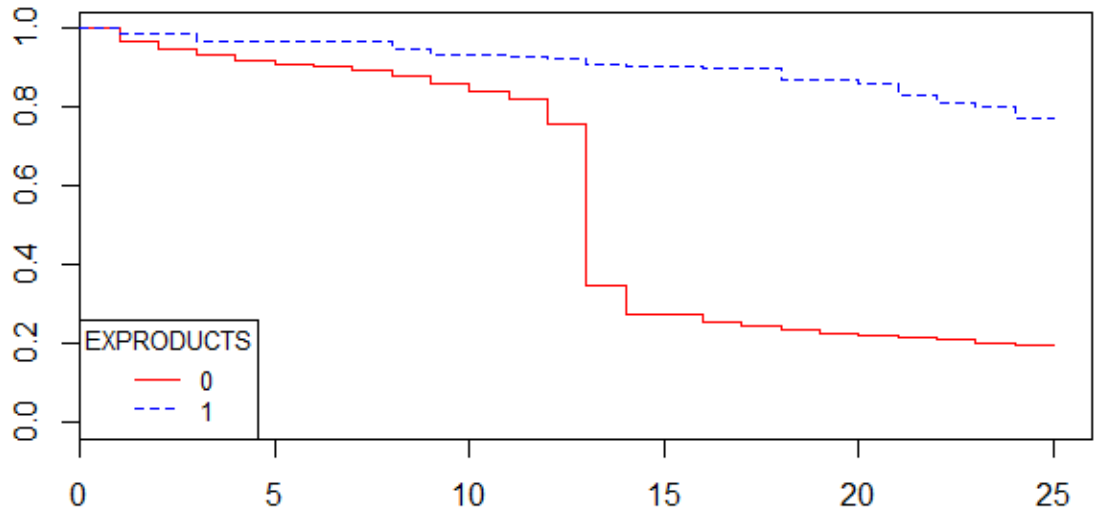


图 4.2 是否办理额外电信服务对应的生存曲线

不难发现，当用户购买了其他的产品时，其最终的生存率比那些没有购买的用户群体高出近 60%。抛开具体的数值，这一现象是符合我们正常认知的，即：当用户购买了同一家电信公司的其他产品，比如购买了某流量包或者开通了某家庭套餐等，其本身会不自觉中更依赖这家公司的服务，或者可以理解成其忠诚度增加了。换个角度，当用户愿意去购

买额外的产品，说明他原本就信任这家公司，而新的购买行为是对其忠诚度的再次强化。这都能解释为什么办理额外的业务的用户群体最终的生存率会高达近 80%。

4.2 Cox 回归分析

Cox 回归模型是由英国统计学家 D.R.Cox(1972)年提出的一种半参数回归模型。该模型以生存结局和生存时间为因变量，可同时分析众多因素对生存期的影响，能分析带有截尾生存时间的资料，且不要求估计资料的生存分布类型。

Cox 回归用公式可以表示为：

$$h(t,X) = h_0(t)exp\left(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_p x_p\right)$$

其中，$\beta_1,\beta_2 \ldots,\beta_p$是自变量的偏回归系数，$X = \left(x_1, x_2 \ldots, x_p\right)$是影响生存时间t的所有变量，$h_0(t)$是基础风险函数(Baseline Survival Function)。若$\beta_i > 0$，则x_i是危险因素；若$\beta_i < 0$，则x_i是保护因素。当X固定时，$\frac{h(t,X)}{h_0(t)}$的值是不变的，与t无关，因此这个模型也成为比例风险模型。我们将$h(t,X)$称为危险函数，其大小定义为危险度。

在描述各个因素的影响时，我们还经常用到相对危险度（Risk Ratio，也叫做风险比 Hazard Ratio），其定义为：暴露组的危险度与对照组危险度之比，用公式可以表示为

$$RR = \frac{h_{(t,X)}}{h_{(t,X^*)}}$$

$$= \frac{h_0(t)exp\left(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_p x_p\right)}{h_0(t)exp\left(\beta_1 x_1^* + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_p x_p\right)}$$

$$= exp\left(\beta_1(x_1 - x_1^*)\right)$$

RR是对单个因素分析的，这里我们以x_1为例。RR的大小意味着暴露组的死亡率是对照组死亡率的多少倍。另外，我们认为，如果RR的值接近 1，那么该因素与死亡没有太大的关系；反之，RR的值偏离 1 的程度越大，其与死亡的关系就越密切。

在 R 中，我们调用 survival 包中的 coxph（）语句对样本进行分析，便可以得到下表：

表 4.3 Cox 回归各变量系数及相关信息

	coef	exp(coef)	se(coef)	z	Pr(>|z|)	
FEE1	-1.064e+00	3.449e-01	1.071e-01	-9.936	< 2e-16	***
EXTIME	9.299e-05	1.000e+00	2.977e-05	3.123	0.00179	**
EXDATA	-7.986e-04	9.992e-01	6.253e-05	-12.771	< 2e-16	***
CHANGE1	-7.035e-01	4.948e-01	1.441e-01	-4.881	1.05e-06	***
CONTRACT1	-1.280e+00	2.781e-01	4.868e-02	-26.289	< 2e-16	***
EXPRODUCTS1	-1.970e+00	1.395e-01	1.705e-01	-11.554	< 2e-16	***
GROUP1	-7.908e-01	4.535e-01	5.004e-02	-15.802	< 2e-16	***

表中第一列 coef 是各个协变量的系数，第二列是对应的相对危险度，最后一列是各个因素的 P 值。

不难发现，除了额外通话时长以外，其他的自变量系数均为负值，其中用户是否购买额外的电信服务产品以及是否签订合约对应的系数绝对值较大。需要注意的是，虽然额外通话时长对应的系数为正值,但是仔细观察我们会发现其数量级为10^{-5},几乎可以记为0。

即使考虑到相当一部分样本的额外通话时长数量级在10^3左右，我们也只能得到一个接近0.1的“系数”。 相反地，虽然额外流量的系数为-7.986×10^{-4}，但是在考虑到其10^3左右的数量级，我们可以将其“系数”视为-0.7986。因此，在以G为单位计量的前提下（$1\mathrm{G} = 1024\mathrm{M}$），额外流量的影响还是相当显著的。这两者对客户流失的影响大小，我们在之后的建模过程中也可以得出。

由于 EXPRODUCTS，CONTRACT，FEE，GROUP 和 CHANGE 三个变量的RR都小于 0.5，因此为了确保用户不流失，在实际操作中，我们应尽可能使这五个变量的值等于 1，其中应最先着手于如何让客户购买额外的电信服务产品以及签订服务合约，这也是为什么我们在日常生活中常常接到电信服务商各种推销电话的原因。

5 基于机器学习的建模与评估

5.1 Logistic 回归

在日常生活中，我们经常需要利用线性回归的方法对某一事物进行预测，比如房价，股票，身高，盈利等，这些变量都有一个共性特征就是连续；而有些情况下，被预测变量可能是二元变量，比如成功或失败，成活或死亡以及本文所研究的流失或不流失等，如果强行使用线性回归的方法可能得不到理想的预测结果，这时我们就需要用逻辑回归（Logistic Regression）的方法进行预测。

下面我们先介绍一下 Logistic 回归将要用到的 Logistic 函数（也称为 sigmoid 函数）：我们定义$S(x) = \frac{1}{1+e^{-x}}$为 logistic 函数，其图像为

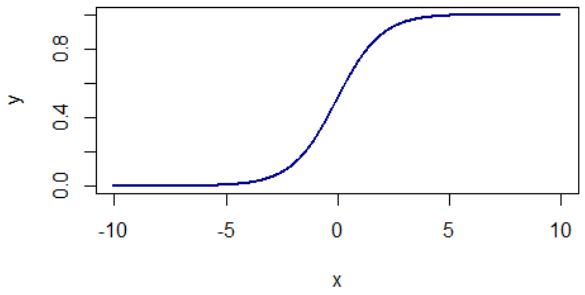


图 5.1 Logistic 函数图像

在一个二元分类问题中，如果我们记事件发生$Y = 1$的概率为P，影响因素为一个p维向量$X = (X_1, X_2 \ldots X_p)$，那么用 logistic 回归表示$P$即为

$$P = \frac{1}{1 + e^{-\theta^T X}}$$

这里θ是一个p维系数向量；同时P也是二元变量Y的期望。

如果我们继续对P做一个 Logit 变换，logistic 回归就可以表示成：

$$logit(P) = ln\left(\frac{P}{1-P}\right) = \theta^T X = \theta_1 X_1 + \theta_2 X_2 + \theta_3 X_3 + \cdots + \theta_p X_p$$

因此，logistic 回归可以看做是对一个 0-1 变量Y的期望做 logit 变换，然后与自变量做线性回归，并对参数采用极大似然估计。

为了得到预测结果，我们通常设置一个阈值P_0，当我们将一个已知的X代入$P=\frac{1}{1+e^{-\theta^T X}}$后，如果$P>P_0$，那么事件$Y$发生，即$Y=1$；反之，$Y=0$。通常，$p_0$取值为 0.5。

5.1.1 建模与检验

在导入 Excel 表格中的样本数据之后，我们还不能直接对数据进行建模，因为此时所有的数据为 factor 类型，我们需要用 as.numeric(as.character())将 EXTIME（额外通话时长）和 EXDATA（额外流量）变成数值型，然后用 glm 作广义线性回归，由于 logistic 回归是建立在二项分布的基础上，所以令 glm 语句中参数 family=binomial（）即可。

在保留交叉检验的训练集上做回归，对所得模型进行 summary 输出：

```
Call:
glm(formula = LOSS ~ ., family = binomial, data = trainset)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8507   0.2815   0.3367   0.3912   2.8644

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.7210182  0.0835934  32.551  < 2e-16 ***
FEE1        -1.2336125  0.1992606  -6.191 5.98e-10 ***
EXTIME      -0.0002514  0.0000659  -3.814 0.000137 ***
EXDATA      -0.0015873  0.0001988  -7.985 1.40e-15 ***
CHANGE1     -1.1853514  0.2956303  -4.010 6.08e-05 ***
CONTRACT1   -1.7484244  0.1084087 -16.128  < 2e-16 ***
EXPRODUCTS1 -2.9844046  0.2632844 -11.335  < 2e-16 ***
GROUP1      -1.7003854  0.1077035 -15.788  < 2e-16 ***
```

上述 P 值表明，我们所用到的这几个变量对客户流失的影响都是显著的，因此不需要去除任何一个变量（当某一个变量对应的 P 值很大时，我们可以考虑将其从模型中移除，因为它对模型几乎没有贡献，这也是为什么我们没有去除 Cox 回归结论中影响“相对”较小的额外通话时长）。同时，根据系数（第一列）绝对值的大小，我们可以确定：一个用户是否办理其它的电信服务，是否签订服务合约以及是否是集团用户对其是否流失有很大影响，这也是为什么我们之前绘图选择 CONTRACT 和 EXPRODUCTS 的原因。

如果我们用公式表示，就有：

$$ln\left(\frac{P}{1-P}\right)=\beta_0+\beta_1 FEE+\beta_2 EXTIME+\beta_3 EXDATA+\beta_4 CHANGE$$
$$+\beta_5 CONTRACT+\beta_6 EXPRODUCTS+\beta_7 GROUP$$

这里β_0=2.7210182 是截距，$\beta_i(i=1\ldots7)$是上图中第一列的数值。注意我们最终是要用这个公式输出的值和我们设置的阈值进行比较，而不是直接采用输出的值。

本文中，为了使结果更具说服性，我们将采用保留交叉验证（Hold-out Cross Validation）和十折交叉验证（10-Fold Cross Validation）两种方法来检验我们模型的好坏。

下面我们用上述的 Logistic 回归函数在保留交叉验证的测试集上进行预测，并与真实值作比较，得到下表（行是预测值，列为真实值）：

表 5.1 Logistic 回归在测试集上的预测结果

	0	1	总计
0	176	71	247
1	151	1095	1246
总计	327	1166	1493

数据表明通过训练集得出的 logistic 回归函数在测试集上的误判率为 0.148693905（设置的阈值$p_0 = 0.5$）。其实我们真正关心的是表中（0,1）格中的数据，即：实际会流失却被我们估计为不会流失的客户的数量，该值所占的百分比为 0.0475552579。

如果我们采用 10 折交叉验证，误判率为 0.149698189（0.0483935743），这里括号里的数据代表实际流失却被估计为不会流失的客户所占的比例，下文中采用类似的格式。

从电信公司决策者的角度来说，5%左右的误判率还算是一个不错的结果。但是从统计学或者数学的角度来说，我们可能接受不了一个模型 15%的误判率。因此我们需要找到一些合适的办法来改进我们的模型。

5.1.2 改进措施

由于模型改进和优化的方法多种多样，而且每种方法都只针对某一种或几种特定的情况。为了找到更适合我们现有模型的优化方法，接下来，我们要用得到的 logistic 函数在保留交叉验证的训练集上作预测，并与真实值作比较，得到下表：

表 5.2 Logistic 回归在训练集上的预测结果

	0	1	总计
0	405	173	578
1	349	2555	2904
总计	754	2728	3482

在训练集上的误判率为 0.149913843，看来通过训练集得到的回归函数在其本身的回归都不够理想。

因为训练误差和测试误差都相当大，说明我们的模型此时是欠拟合(Underfitting)的[12]。如果想要得到更好的预测模型，我们需要增加现有模型中的自变量，即增加模型的复杂程度，比如用户月均主叫次数，月均短信条数等。当然，增加自变量的数量也是有限度的，当自变量数量太多时，就会出现过拟合（Overfitting）的情况，即：模型可以“完美”预测训练集的数据，却不具有泛化问题的能力（测试集表现很差）。由于数据来源的限制，我们无法得到原始数据表里这 4975 名用户的其他信息，所以我们这里也只能从理论上分析改进方法。

5.2 决策树

决策树（Decision Tree）既是一种在机器学习中经常用到的方法，也是一种在日常生活中我们潜意识里经常用到的方法。它是一种基于树的模型，由决策节点，决策分支和叶子三个部分组成。实际上，各种决策树都是根据某种规则对训练数据进行递归式的分类，并最终根据终止条件结束算法。根据每个节点最优值的确定标准，决策树算法可以大致分为 ID3，C4.5 和 CART[13]（Classification and Regression Tree）。

5.2.1 建模与检验

我们先调用 R 里 C5.0 包对保留交叉验证的训练集（70%）中的数据进行建模。

```
treecontrol<-C5.0Control(CF=0.25,winnow=T,noGlobalPruning=F,minCases =20)
treemodel <- C5.0(LOSS~.,data=trainset,rules=F,control =treecontrol)
```

这里 CF 是剪枝的置信度，winnow 指在建模之前是否对模型进行变量筛选，noGlobalPruning=F 意味着我们允许 R 对我们的 C5.0 决策树进行剪枝，minCases 可以理解成最终每个分类里样本个数的最小值。plot（treemodel）绘图之后我们便可以得到如下的决策树：

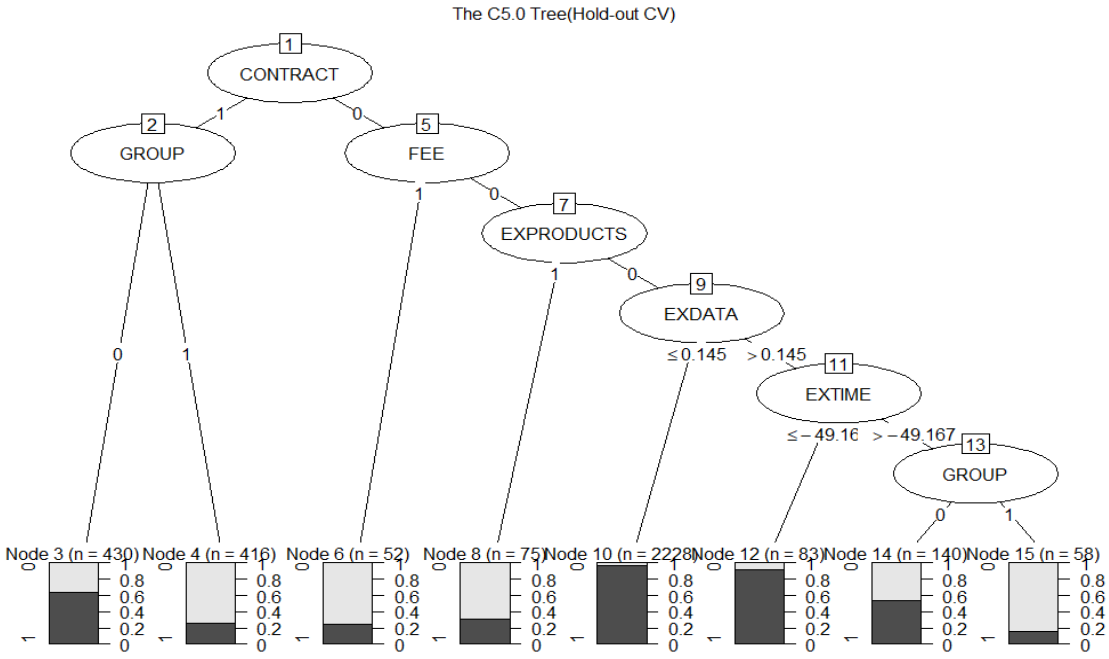


图 5.2 C5.0 决策树

在测试集上使用该决策树，得到的预测结果和真实数据对比如下（行是预测值，列是真实值）：

表 5.3 C5.0 决策树在测试集上的预测结果

	0	1	总计
0	192	59	251
1	135	1107	1242
总计	327	1166	1493

因此，预测的误判率为 0.129939719（0.0395177495）。

在训练集上的误判率为 0.132969558。

接下来我们试着用 CART 算法解决这个问题。下面是建模的语句

```
cartmodel_holdout<-rpart(LOSS~.,data=trainset,method = "class")
```

调用 rpart.plot（）之后可以得到下图

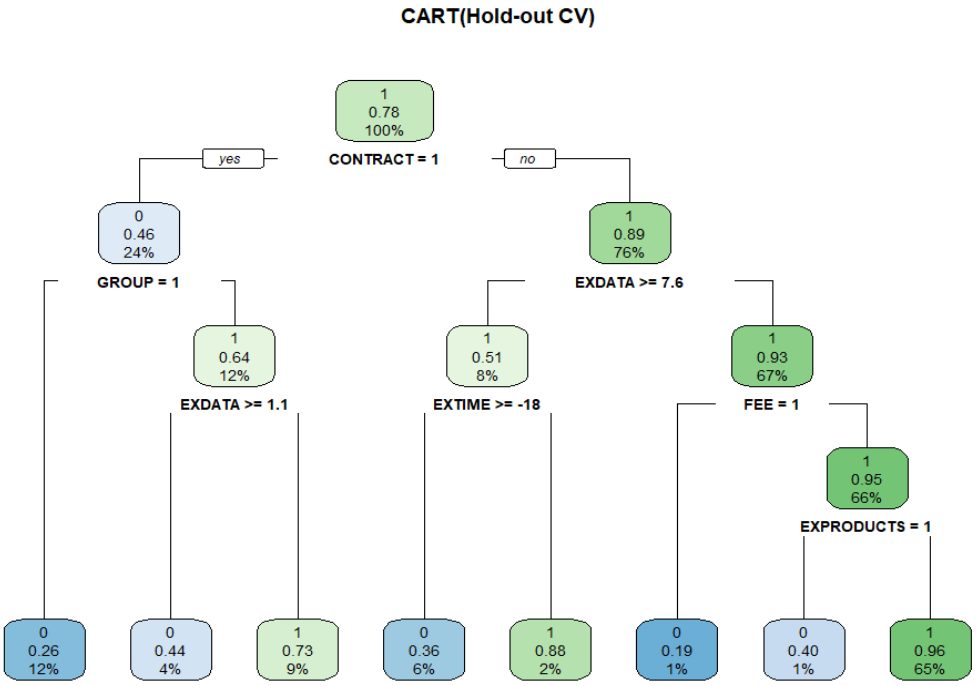


图 5.3 CART 树

注：(1)每一个左侧分支代表 YES，右侧分支代表 NO；

(2)以第二行右侧的分支节点为例，1 代表当不满足 CONTRACT=1 时，我们预测该用户流失，0.89 是该预测的准确率，76%意味着训练集中有 76%的数据落在“CONTRACT=1”为 FALSE 的集合内。

此时测试集和训练集的误判率分别为 0.131949096 和 0.129236071。

我们把 10 折交叉检验的结果和上述结果汇总在一起，列出下表

表 5.4 决策树结果误判率结果汇总

验证方法	保留交叉验证		十折交叉验证	
集合	测试集	训练集	测试集	训练集
C5.0	0.129939719 （0.0395177495）	0.132969558	0.132931727 （0.0516064257）	0.130779540
CART	0.138647019 （0.079035499）	0.129236071	0.140763052 （0.0775100402）	0.130332812

注：括号里的数字是实际流失却被预测为不会流失的人数所占的百分比。

综上，C5.0 决策树的误判率在 13%左右，CART 误判率相对高了一个百分点；同时，括号内的数值表明，由 C5.0 决策树得到的预判可以更有效地防止客户流失，进而提高电信公司的盈利。

5.2.2 改进措施

通过 5.1.2 中的分析，我们知道可以通过增加自变量的数量来改善我们的模型。和逻辑回归中筛选自变量类似的是：为了避免决策树中的自变量类型太多而导致我们的模型出现过拟合的情况，我们可以调用某些语句对树进行“剪枝”操作，进而使我们的预测模型更为强健。

6 总结与展望

对客户流失问题的研究是在市场需求的刺激下发展起来的，其研究成果对于降低企业运营成本并提高盈利，具有重要的指导作用。本文主要运用了生存分析，Logistic 回归和决策树三种方法对该问题进行了研究，其中生存分析着重于探究各个因素的影响，logistic 回归和决策树重点则在于建立合适的模型进行预测。

总结起来，我们大致可以得到以下两点结论：

（1）一个用户购买额外的电信服务产品的数量以及是否签订合约对其将来是否流失具有很大的影响，因此，为了避免用户流失，电信公司需要制定足够合理的营销策略来吸引老用户购买或签订新的产品或合约。

（2）根据保留交叉验证以及十折交叉验证的结果，Logistic 回归和两种决策树（C5.0 和 CART）在预测时的表现都还可以接受：三个模型在测试集上误判率均在 13%-15%左右，其中更容易造成电信公司客户流失的误判（即实际会流失而模型预测为不会流失）所占比例在 4%-8%左右。根据本文所得出的结果，C5.0 决策树的表现更好一些。虽然经典的 Logistic 回归在测试集上的误判率高达 15%，但是（0,1）对应的值只有 5%左右，从电信公司的角度来说这个结果还是不错的。

虽然关于客户流失问题的研究已经较为成熟，但是由于科技的进步和市场的变化，我们仍需要对现有的模型不断地进行调整优化。就本文整个研究过程而言，在未来的研究中可以针对以下几点进行：

（1）数据的采集应更加全面

通过对比训练集和测试集上的误判率可以得出，我们现有的三个模型复杂度均不够，即因变量的数量太少，可供我们筛选的条件太少，因此预测的结果不够精确。条件允许的情况下，我们可以通过增加诸如用户月主叫次数，月短信条数，日间上网流量，夜间上网流量等新的统计量，来提高模型预测的准确率。

（2）对现有模型优化的探究

在 Logistic 回归中，我们可以通过尝试用其他的算法来计算每一项的系数，比如牛顿迭代法（Newton’s Method），然后在新的系数下计算测试集上的误判率；在决策树中，我们可以增加对树各项参数的限制或者改变现有参数的值，比如树的深度，节点个数，根节点的样本数量以及剪枝的置信度等，然后观察新的树是否有更好的表现。

（3）客户流失原因的分析与评估

客户流失不是偶然发生的，它是一个动态过程，可以借鉴系统动力学的思想考察对客户流失有重要影响的因素，分析其影响客户流失的机理，这是客户流失原因分析与评估的研究方向之一[14]。同时，目前大多数流失原因的分析都是基于专业领域的各种知识，从某种角度来说，通过专业知识得出的结论更具有说服力，但是我们也许应该适当地把目光放在客户身上，特别是那些有可能流失的客户，如果能将他们的反馈和专业知识有效地结合，也许我们可以得到更理想、更准确的分析结果。

参考文献

[1]舒文琼.BMC：解决运营商客户流失之困[J].通信世界,2008(12):74.
[2]林向阳. 基于数据挖掘的移动大客户流失分析.[D].清华大学,2009.
[3] 迟准.电信运营企业客户流失预测与评价研究[D].哈尔滨工程大学,2013.
[4] 刘勇.中国电信业流失客户赢回策略研究[D].华中科技大学,2007.
[5] 单友成. CRM 中模糊数据挖掘及客户生命周期价值与客户满意度研究[D].天津大学,2009.
[6] 吕昀卿.商业银行客户细分及应用研究[D].对外经济贸易大学,2006.
[7] Lawless J F Statistical models and Methods for Lifetime Data[M]. John Wiley&Sons, Inc., New York,2002.
[8] 环梅.基于生存分析的信号交叉口非机动车穿越行为研究[D].北京交通大学,2014.
[9] T. M. Mitchell. Machine Learning[M].New York:McGraw‐Hill, 1997.
[10] 陈慧灵. 面向智能决策问题的机器学习方法研究[D].吉林大学,2012.
[11] 刘朝华.基于客户价值的客户分类模型研究[D].华中科技大学,2008.
[12] Joram Soch,John-Dylan Haynes,Carsten Allefeld. How to avoid mismodelling in GLM-based fMRI data analysis: cross-validated Bayesian model selection[J]. NeuroImage,2016,141.
[13] 张桡,曹健.面向大数据分析的决策树算法[J].计算机科学,2016,43(S1):374-379+383.
[14] 于小兵,曹杰,巩在武.客户流失问题研究综述[J].计算机集成制造系统,2012,18(10):2253-2263.
[15]Reicheld F F, Sasser W E. Zero defection: Quality comes to services[J]. Harvard Business Review,1990,429(9)：105-110.
[16]Drew, James H, D. R. Mani, Andrew L. Betz and Piew Datta, Targeting Customers with Statistical and Data-mining Techniques[J]. Journal of Service Research,2001,3(3):205-219
[17] LOUISA C 2002,21 (2):Data mining and causal modeling of customer [J]. Tele-communication System s, 381-394.
[18]朱娅婷. 中国电信 S 分公司宽带客户流失成因及对策研究[D].苏州大学,2015.
[19]柳兰屏,曾煌.移动通信客户流失分析方法[J].移动通信,2003(4):97-99.

致谢

值此论文完成之际，我衷心感谢我的论文指导老师李锋。作为一名应用数学的学生，我在完成本篇论文期间遇到了很多疑惑和困难，但是在李老师的耐心指导和帮助下，所有问题都被一一化解，李老师的专业热情也促使我更加想要在以后的学习生活中学好统计学这门课程。同时，我也要感谢大学期间其他的任课老师们，您们这四年孜孜不倦的教导使我收获了很多专业知识，也提升了我自主学习的能力，希望各位老师在以后的工作生活中身体健康，万事如意！

附录

```
library(xlsx)
A<-read.xlsx2(file="C:\\Users\\GuoLY\\Desktop\\毕业论文\\CUSTOMERLOSS.xlsx",sheetName=1)
A$EXTIME<-as.numeric(as.character(A$EXTIME))
A$EXDATA<-as.numeric(as.character(A$EXDATA))
A$DURATION<-as.integer(as.character(A$DURATION))
table(A$FEE)
table(A$CHANGE)
table(A$CONTRACT)
table(A$EXPRODUCTS)
table(A$GROUP)
table(A$LOSS)
hist(A$EXTIME,breaks = seq (-3000,5000,100))
hist(A$EXDATA,breaks = seq (-2200,2600,100))
summary(A$EXTIME)
summary(A$EXDATA)
hist(A$DURATION)
summary(A$DURATION)
table(LOSS,FEE)
table(LOSS,CHANGE)
table(LOSS,CONTRACT)
table(LOSS,GROUP)
table(LOSS,PRODUCTS)
table(LOSS,EXPRODUCTS)
library(ggplot2)
ggplot(A,aes(x=FEE,fill=LOSS))+geom_histogram(stat="count",position="dodge")
ggplot(A,aes(x=CONTRACT,fill=LOSS))+geom_histogram(stat="count",position="dodge")
ggplot(A,aes(x=CHANGE,fill=LOSS))+geom_histogram(stat="count",position="dodge")
ggplot(A,aes(x=EXPRODUCTS,fill=LOSS))+geom_histogram(stat="count",position="dodge")
ggplot(A,aes(x=GROUP,fill=LOSS))+geom_histogram(stat="count",position="dodge")
ggplot(A,aes(x=EXPRODUCTS,fill=LOSS))+geom_histogram(stat="count",position="dodge")
library(xlsx)
B<-read.xlsx2(file="C:\\Users\\GuoLY\\Desktop\\毕业论文\\CUSTOMERLOSS.xlsx",sheetIndex=2)
C<-read.xlsx2(file="C:\\Users\\GuoLY\\Desktop\\毕业论文\\CUSTOMERLOSS.xlsx",sheetName=2)
hist(as.numeric(as.character(B$EXTIME)),main = "Histogram of EXTIME(LOSS=0)",xlab = "EXTIME",col = "lightblue")
```

```
hist(as.numeric(as.character(B$EXDATA)),main = "Histogram of EXDATA(LOSS=0)
",xlab = "EXDATA",col = "lightblue")
hist(as.numeric(as.character(C$EXTIME)),main = "Histogram of EXTIME(LOSS=1)
",xlab = "EXTIME",col = "orange")
hist(as.numeric(as.character(C$EXDATA)),main = "Histogram of EXDATA(LOSS=1)
",xlab = "EXDATA",col = "orange")
library(survival)
surv<-survfit(Surv(A$DURATION,A$LOSS==1)~1)
summary(surv)
plot(surv)
surv_EXPRODUCTS<-survfit(Surv(A$DURATION,A$LOSS==1)~A$EXPRODUCTS)
summary（surv_EXPRODUCTS）
opar<-par(no.readonly = T)
plot(surv_EXPRODUCTS,lty=c(1,2),col=c("red","blue"))
legend("bottomleft",c("0","1"),title="EXPRODUCTS",lty=c(1,2),col=c("red","blu
e"))
par(opar)
coxfit<-coxph(Surv(DURATION,as.numeric(LOSS)==2)~
                FEE+EXTIME+EXDATA+CHANGE+CONTRACT+EXPRODUCTS+GROUP,data=A)
summary(coxfit)
P<-read.xlsx2(file="C:\\Users\\GuoLY\\Desktop\\      毕      业      论      文
\\CUSTOMERLOSS01.xlsx",sheetName=1)
P$EXDATA<-as.numeric(as.character(P$EXDATA))
P$DURATION<-as.integer(as.character(P$DURATION))
  l<-glm(LOSS~.,data=P,family=binomial())
summary(l)
m=4975
trainindex<-sample(1:m,round(0.7*m))
trainset<-P[trainindex,]
testset<-P[-trainindex.]
Lguess<-predict(l,type=" response" ,newdata=testset)
guess<-ifelse(Lguess>0.5,1,0)
table(guess,testset$LOSS)
Lguesstrain<-predict(L,type="response",newdata=trainset)
guesstrain<-ifelse(Lguesstrain>0.5,1,0)
table(guesstrain,trainset$LOSS)
for(i in 1:10)
{foldtest<-P[folds[[i]],]
 foldtrain<-P[-folds[[i]],]
 foldL<-glm(LOSS~.,data=foldtrain,family = binomial)
 foldpre<-predict(foldL,type = "response",newdata = foldtest)
 foldpre<-ifelse(foldpre>0.5,1,0)
 print(table(foldpre,foldtest$LOSS))}
#C50+holdout
install.packages("C50")#错误率 0.129939719,训练集 0.132969558
```

```
library(C50)
treecontrol<-C5.0Control(CF=0.25,winnow=T,noGlobalPruning=F,minCases =20)
treemodel <- C5.0(LOSS~.,data=trainset,rules=F,control =treecontrol)
treepre<-predict(treemodel,testset)
table(treepre,testset$LOSS)
treepre_train<-predict(treemodel,trainset)
table(treepre_train,trainset$LOSS)
table(treepre)
C5imp(treemodel)
plot(treemodel,main="The C5.0 Tree(Hold-out CV)",font.main=4)
#C50 + 10-fold
library(C50)
install.packages("rpart.plot")
library(rpart)
library(rpart.plot)
require(caret)
#10-fold method+ CART(错误率 0.140763052) trainset  0.130332812
folds<-createFolds(y=P$LOSS,k=10)
sum_error=0
sum_error1=0
sum_errortrain=0
for(i in 1:10){
  foldtest<-P[folds[[i]],]
  foldtrain<-P[-folds[[i]],]
  cartmodel_fold<-rpart(LOSS~.,data=foldtrain,method="class")
  cartpre<-predict(cartmodel_fold,foldtest)
  cartpre<-ifelse(cartpre[,2]>0.5,1,0)
  cartpre_train<-predict(cartmodel_fold,foldtrain)
  cartpre_train<-ifelse(cartpre_train[,2]>0.5,1,0)
  sum_error<-sum_error+sum(ifelse(foldtest$LOSS==cartpre,0,1))
  sum_error1<-sum_error1+sum(table(cartpre,foldtest$LOSS)[1,2])
  sum_errortrain<-
sum_errortrain+sum(ifelse(foldtrain$LOSS==cartpre_train,0,1))
  }
print(sum_error)
print(sum_error1)
print(sum_errortrain)
?rpart
library(rpart)
library(rpart.plot)
#hold-out+ cart
cartmodel_holdout<-rpart(LOSS~.,data=trainset,method = "class")
rpart.plot(cartmodel_holdout,main="CART(Hold-out CV)",cex=0.7)
#prediction(错误率 0.138647019) trainset (0.129236071)
cartpre_holdout<-predict(cartmodel_holdout,testset)
```

```
cartpre_holdout<-ifelse(cartpre_holdout[,2]>0.5,1,0)
table(cartpre_holdout,testset$LOSS)
cartpre_holdout
table(cartpre_holdout)
cartpre_holdouttrain<-predict(cartmodel_holdout,trainset)
cartpre_holdouttrain<-ifelse(cartpre_holdouttrain[,2]>0.5,1,0)
table(cartpre_holdouttrain,trainset$LOSS)
save.image()
savehistory()
q()
```