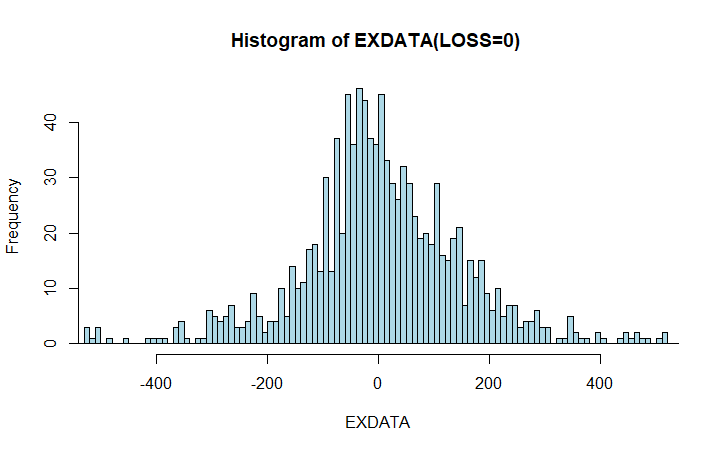
2.3.3 机器学习与数据挖掘

机器学习与数据挖掘经常被一起提起，在很多场合它们甚至被认为是同样的概念。事实上，数据挖掘可以视为机器学习，数据库，统计学以及各种学科的交叉，它主要利用机器学习的各种方法来分析数据，利用数据库的技术来管理海量数据。由于统计学界往往醉心于理论的优美而忽视实际的效用，因此，统计学界提供的很多技术通常都要在机器学习界进一步研究，变成有效的机器学习算法之后 才能再进入数据挖掘领域。从这个意义上说，统计学主要是通过机器学习来对数据挖掘发挥影响，而机器学习和数据库则是数据挖掘的两大支撑技术。

从数据分析的角度来看，绝大多数数据挖掘技术都来自机器学习领域。但数据挖掘并不只是数据分析技术的简单应用。一个重要的区别是，传统的机器学习研究并不把海量数据作为处理对象，很多技术是为处理中小规模数据设计的，如果直接把这些技术用于海量数据，效果可能很差，甚至可能用不起来。因此，数据挖掘界必须对这些技术进行专门的、不简单的改造。

另一方面，作为一个独立的学科领域，必然会有一些相对“独特”的东西。对数据挖掘来说，这就是关联分析。简单地说，关联分析就是希望从数据中找出“买尿布的人很可能会买啤酒”这样看起来匪夷所思但可能很有意义的模式h。在R. Agrawal等人首先对关联规则挖掘进行研究之后，大批学者投身到这方面的研究中并产生了很多成果，代表性工作有R. Agrawal和R. Srikant的Apriori算法以及J. Han等人的FP-Growth算法等[11][13]。

另外，对于通话时长，有超过一半的流失用户都存在通话时长“过剩”的情况，而很多没有流失的用户通话时长均大于套餐时长，其中大多数额外时长落在[0,500]的区间内。在流量方面，绝大多数流失用户也都有“过剩”的现象，比较有意思的是，观察期过后仍然使用该电信公司手机号的用户的流量使用呈现明显的对称分布，且集中在[-500,500]之间。统计学知识告诉我们这很可能是一个正态分布，为了进一步验证，我们只取[-500,500]之间的分布，得到图3.8

目录

1 引言

1.1 问题背景

1.2 研究目的及意义

1.3 国内外研究情况

2 于机器学习的客户流失研究综述

2.1 客户流失的定义

2.2 客户流失管理

2.3 生存分析简介

2.4 机器学习简介

2.4.1 机器学习定义

2.4.2 机器学习方法分类

3 于生存分析与机器学习对手机客户流失问题的分析与建模

3.1 数据准备

3.1.1 数据来源

3.1.2 变量引入

3.1.3 数据预处理

3.2基于生存分析的研究

3.2.1 生存过程的描述

3.2.2 生存过程的比较

3.2.3 Cox回归分析

3.2.3.1 Cox回归简介

3.2.3.2 分析结果

3.3 基于逻辑回归的建模与评估

3.3.1 简介

3.3.2 建模与检验

3.3.3 改进措施

3.4 基于决策树的建模与评估

3.4.1 简介

3.4.2 建模与检验

3.4.3 改进措施

4 总结与展望

参考文献

致谢

附录

1 引言......................................................................1

1.1 问题背景..............................................................1

1.2 研究目的及意义........................................................1

1.3 国内外研究情况........................................................1

2 关于手机客户流失问题研究综述..............................................2

2.1 客户流失的定义........................................................2

2.2 客户流失管理..........................................................2

2.3 生存分析简介..........................................................2

2.4 机器学习简介..........................................................3

3 数据准备..................................................................3

3.1 数据来源..............................................................3

3.2 变量引入..............................................................3

3.3 数据预处理............................................................4

4 基于生存分析的研究........................................................8

4.1 生存过程的描述........................................................8

4.2 生存过程的比较........................................................9

4.3 Cox 回归分析.........................................................10

5 基于机器学习的建模与评估.................................................11

5.1 Logistic 回归........................................................11

5.1.1 建模与检验.....................................................12

5.1.2 改进措施.......................................................13

5.2 决策树...............................................................14

5.2.1 建模与检验.....................................................14

5.2.2 改进措施.......................................................16

6 总结与展望...............................................................16

参考文献...................................................................18

致谢.......................................................................19

附录.......................................................................20