基于逻辑回归算法，搭建客户流失预测模型

◆ 孙 琳

摘要：中心的宽带数据每年都以20%的速度增长，但是固话业务虽然用户是增长趋势，但是收入确实萎缩的。

面临经营压力，什么原因导致固话业务收入的减少，油区客户拨打固话的习惯又是什么？什么样的客户不在使用固话业务？本论文对客户进行细分，利用逻辑回归模型实现固话客户流失预测分析。

关键词：固话业务；萎缩；客户流失

一、现状分析

中心的宽带数据每年都以 20% 的速度增长，但是固话业务虽然用户是增长趋势，但是收入确实萎缩的。面临经营压力，通过大数据技术，对客户进行细分，利用逻辑回归模型实现固话客户流失预测分析。

二、逻辑回归算法

y=f(x)，表明自变量 x 与因变量 y 的关系，通过 N 多个因素判断事物是否发生。简单说：医生治病时的望、闻、问、切获取自变量 x，即特征数据，判断是否生病（y）。

三、利用逻辑回归建立客户流失模型

业界普遍都是采用逻辑回归算法来建立模型。要想做好客户流失分析，最重要的是明确业务问题的定义和如何运用挖掘结果来对客户进行挽留。

1. 明确业务问题定义。数据挖掘就是个不断尝试的过程，首先要明确需求，就是需求分析，只有明确了业务问题才能避免多走弯路，浪费人力物力。

对于客户流失预测来说，一般要明确这几个问题：

—什么叫做流失？什么叫做正常？（严格定义好 0 和 1）

—要分析哪些客户？

2. 变量选取、数据探索和多次建模

分为如下几类：

—客户基本信息（年龄、性别、在网时长、当前状态 ..... ）

—客户账单信息（账单金额、优惠金额、明细账单金额 ..... ）

—客户缴费信息（缴费次数、缴费金额、欠费次数、欠费金额 ..... ）

—客户通话信息（通话次数、通话时长、短信次数、呼转次数、漫游次数 ..... ）

—客户联络信息（投诉次数、抱怨次数 ..... ）

3. 利用 SPSS 搭建客户流失模型

1）固话客户拨打习惯分析

数据范围：提取客户局内、市话、长话次数和时长特征值，共 70 多万条数据。

聚类方法：利用 K– 均值聚类方法

聚类结果：通过描述性统计，发现长途拨打时长、次数、局内拨打时长、次数、市话拨打时长、次数与总收都有比较密切的关系。

第一类客户。无论是局内、市话、长话，其拨打次数都很少，整体的贡献程度很低；

第二类客户。局内拨打次数很多，市话拨打次数也相对较多。

第三类客户。只有局内拨打记录，且次数是五类中最高的。

第四类客户。长话拨打次数最多，而且长途收入最高。

第五类客户。市话和长话的拨打次数一般，但总的拨打次数与第四类客户的总拨打次数相当。

2）利用 SPSS 搭建客户流失模型。利用逻辑回归建立流失模型。

模型分析：市话时长每上一个等级，离网的风险要增加 25% 左右，同理长途拨打次数每上一个等级，其离网的风险降低 29%。局内电话及市话使用越多，其离网的可能性越高，长途使用越多，其离网的风险越低。 之后我们对即将离网的客户，进行客户价值聚类发现 78% 客户为中端客户。

建议：可以推出一些市话内的套餐，增加客户的粘性。

采取措施：

①基于离网客户测算其平均使用时长，以及基于市话使用的分段。

②根据市话使用的分段，挖掘目标客户设计套餐的时长及资费。

③计算套餐的拉新量，以及套餐推出预测后半年后的收入。

结论

基于大数据的客户关系管理体系在设计伊始，“以市场为中心，以客户需求为导向”的目标就非常的清晰而坚定。对客户流失分析，可以使我们在第一时间找到客户流失的模式，比如说客户的活跃度降低，或者客户购买产品的品类发生了变化。洞察客户的心理，才能使我们获得更大的商机。

（作者单位：天津市大港油田信息中心）