工程技术与管理·第 04 卷·第 05 期·2020 年 05 月
DOI: https://doi.org/10.26549/gcjsygl.v4i5.4189

研究性文章
Article

Analysis of the Loss of Telecommunications Users in the Framework of Big Data Systems

Liang Feng

Taiyuan University of Technology, Taiyuan, Shanxi, 030000, China

Abstract

At present, telecom operators are facing fierce market competition. For telecom operators, customers are the life, how to maintain existing customers is the top priority of enterprise customer management. Therefore, the more customers the telecommunications operator has, the greater the initial investment as the main cost, and the greater the profit of the enterprise. The significance of customer resources to telecom operators is self-evident. The competition between telecom operators is actually the competition for customer resources. The paper starts with the relationship between user characteristics and churn, and gives suggestions for increasing user stickiness and preventing churn for reference.

Keywords

big data system; customer architecture; analysis

关于大数据系统构架中电信用户流失的分析

冯亮

太原理工大学，中国·山西 太原 030000

摘　要

目前，电信运营商面临着激烈的市场竞争。对电信运营商来说客户即生命，如何保持现有客户是企业客户管理的重中之重。因此，电信运营商拥有的客户越多，作为主要成本的前期投资就会越大，企业的利润也就越大。客户资源对电信运营商的意义不言而喻，电信运营商之间的竞争实际上就是对客户资源的竞争。论文从用户特征与流失关系入手，针对性给出增加用户黏性、预防流失的建议，以供参考。

关键词

大数据系统；客户构架；分析

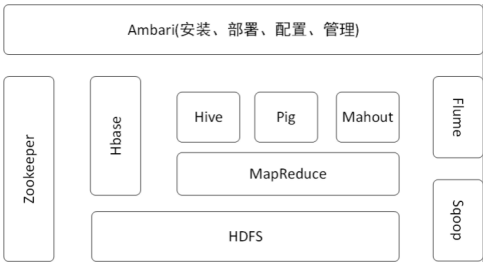
1 引言

当今电信市场竞争激烈运营商每月客户流失率在1%~3%，挽留将要流失客户、降低客户流失率是近年来热门的研究领域[1]。而数据挖掘技术是解决这一问题的有效途径，论文对数据集进行数据挖掘与分析，深入了解电信客户流失的关键，以对该类客户的行为特性进行预警分析，采取针对性的措施改善客户关系，避免客户流失或者挽留客户。[2,3] 文中数据是在CCF大数据与计算智能大赛官网寻找，来源于Kaggle平台。数据集的大小为7043行，22列，并且存在缺失。

Hadoop平台对处理大数据本身具有很显著的优点，首先它具有很高的可靠性，Hadoop中HDFS分布式文件系统采用了备份恢复机制，MapReduce中的任务采用了监控机制，这就保证了分布式处理的可靠性；其次它具有很好的高扩展性，Hadoop是在可用的计算机集群间进行数据的分配的，也是在集群中分布完成计算任务的，这些集群族能扩展到数以千计的节点中，Hadoop能可靠的存储和处理数据。不管在存储上还是计算上，可扩展性都是Hadoop的设计根本所在；并且它具有高效性，Hadoop的高效性表现在Hadoop能够在节点之间进行动态的移动数据，同时能保证各个节点的数据动态平衡，这就使得Hadoop在处理数据时速度非常快。这种方式为高效处理海量数据做好了基础准备。Hadoop可以运行在廉价PC上，采用自动保存数据的多个副本方式，并能自动为失败的任务进行重新配置。随着Hadoop生态系统的成长，越来越多的新项目对Hadoop是很好的补充或提供一些更高层的抽象。

Hadoop的生态图如下：

工程技术与管理·第 04 卷·第 05 期·2020 年 05 月
DOI: https://doi.org/10.26549/gcjsygl.v4i5.4189



（1）HDFS：分布式文件系统，可以对数据进行存储。

（2）MapReduce：分布式数据处理模型和执行环境，可以对数据进行处理操作。

（3）ZooKeeper：在分布式系统中如何就某个值（决议）达成一致，是一个十分重要的基础问题。ZooKeeper 作为一个分布式的服务框架，解决了分布式计算中的一致性问题。在此基础上，ZooKeeper 可用于处理分布式应用中经常遇到的一些数据管理问题，如统一命名服务、状态同步服务、集群管理、分布式应用配置项的管理等。ZooKeeper 常作为其他 Hadoop 相关项目的主要组件，发挥着越来越重要的作用。

（4）Hbase：Hbase 是一个在 HDFS 上开发的面向列的分布式数据库。如果需要实时地随机访问超大规模数据集，我们就可以使用 Hbase 这一 Hadoop 应用。

（5）Pig：运行在 Hadoop 上，是对大型数据集进行分析和评估的平台。它简化了使用 Hadoop 进行数据分析的要求，提供了一个高层次的、面向领域的抽象语言：PigLatin。通过 Pig Latin，数据工程师可以将复杂且相互关联的数据分析任务编码为 Pig 操作上的数据流脚本，通过将该脚本转换为 MapReduce 任务链，在 Hadoop 上执行。和 Hive一样，Pig 降低了对大型数据集进行分析和评估的门槛。

（6）Hive：是 Hadoop 中的一个重要子项目，最早由 Facebook 设计，是建立在 Hadoop 基础上的数据仓库架构，它为数据仓库的管理提供了许多功能，包括：数据 ETL（抽取、转换和加载）工具、数据存储管理和大型数据集的查询和分析能力。Hive 提供的是一种结构化数据的机制，定义了类似于传统关系数据库中的类 SQL 语言。

（7）Mahout：起源于 2008 年，最初是 Apache Lucent 的子项目，它在极短的时间内取得了长足的发展，现在是 Apache 的顶级项目。Mahout 的主要目标是创建一些可扩展的机器学习领域经典算法的实现，旨在帮助开发人员更加方便快捷地创建智能应用程序。Mahout现在已经包含了聚类、分类、推荐引擎（协同过滤）和频繁集挖掘等广泛使用的数据挖掘方法。除了算法，Mahout 还包含数据的输入 / 输出工具、与其他存储系统（如数据库、MongoDB 或 Cassandra）集成等数据挖掘支持架构。

（8）Hume：Flume 是 Cloudera 开发维护的分布式、可靠、高可用的日志收集系统。它将数据从产生、传输、处理并最终写入目标的路径的过程抽象为数据流，在具体的数据流中，数据源支持在 Flume 中定制数据发送方，从而支持收集各种不同协议数据。同时，Flume 数据流提供对日志数据进行简单处理的能力，如过滤、格式转换等。此外，Flume 还具有能够将日志写往各种数据目标（可定制）的能力。总的来说，Flume 是一个可扩展、适合复杂环境的海量日志收集系统。

（9）Sqoop：是 SQL−to−Hadoop 的缩写，是 Hadoop 的周边工具，它的主要作用是在结构化数据存储与 Hadoop 之间进行数据交换。Sqoop 可以将一个关系型数据库（例如 MySQL、Oracle、PostgreSQL 等）中的数据导入 Hadoop 的 HDFS、Hive 中，也可以将 HDFS、Hive 中的数据导入关系型数据库中。Sqoop 充分利用了 Hadoop 的优点，整个数据导入导出过程都是用 MapReduce 实现并行化，同时，该过程中的大部分步骤自动执行，非常方便。

2 具体数据分析

2.1 实验环境搭建

Hadoop3.2.0，Hive3.1.2，Sqoop1.4.7，Spark2.4.4。

2.2 数据预处理

（1）导入数据集，并查看数据及信息、大小。

（2）检查各列、各字段数据类型、字段内容和数量，发现“TotalCharges”（总消费额）列有 11 个用户数据缺失，将其数据类型强制转换为浮点型，并将缺失用户数据填充为“NaN”。

（3）经过观察，发现这 11 个用户‘tenure’（入网时长）为 0 个月，推测是当月新入网用户。根据一般经验，用户即使在注册的当月流失，也需缴纳当月费用。因此将这 11 个用户入网时长“tensure”改为 1，将总消费额填充为月消费额，符合实际情况。

（4）将处理完的数据保存为新的数据集。

2.3 使用 Hive 数据分析

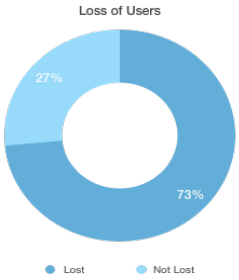
将数据加载到 Hive 中

（1）将预处理后的新数据集上传到 HDFS 中。

（2）在 Hive 中创建一个数据库来加载 HDFS 中的数据

2.4 分析用户各属性及流失率的关系

2.4.1 分析流失用户数量和占比（见图 1）



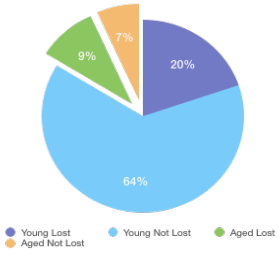
LOSS OF USERS	
Lost	5,174
Not Lost	1,869
TOTAL LOSS	7,043

图 1 流失用户数量和占比

结论：属于不平衡数据集，流失用户占比达 26.54%。

2.4.2 用户属性分析

按照年龄分析用户流失比例，如图 2 所示。

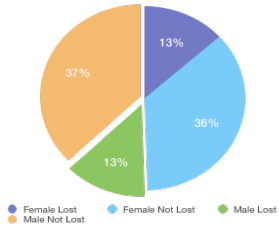


AGE	
Young Lost	1,393
Young Not Lost	4,508
Aged Lost	666
Aged Not Lost	476

图 2 分析用户流失比例图（按照年龄）

结论：年老用户流失率占显著高于年轻用户。

按照性别分析用户流失比例，如图 3 所示。

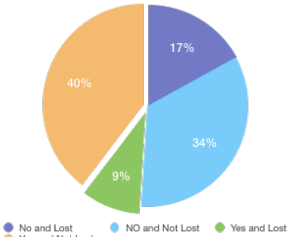


GENDER	
Female Lost	939
Female Not Lost	2,549
Male Lost	930
Male Not Lost	2,625

图 3 分析用户流失比例（按照性别）

结论：男性与女性用户之间的流失情况基本没有差异。

按照婚否分析用户流失比例，如图 4 所示。

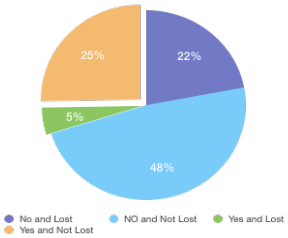


MARRIED OR NOT	
No and Lost	1,197
NO and Not Lost	2,394
Yes and Lost	662
Yes and Not Lost	2,790

图 4 分析用户流失比例（按照婚否）

结论：在所有数据中未婚与已婚人数基本持平，但未婚中流失人数比已婚中的流失人数高出了快一倍。

按照是否有家属分析用户流失比例，如图 5 所示。



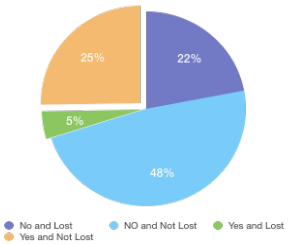
HAVING CHILDREN OR NOT	
No and Lost	1,549
NO and Not Lost	3,391
Yes and Lost	324
Yes and Not Lost	1,779

图 5 分析用户流失比例（按照是否有家属）

结论：有家属的用户流失占比低于无家属用户。

2.4.3 服务属性分析

按照有多条线路分析用户流失比例，如图 6 所示。

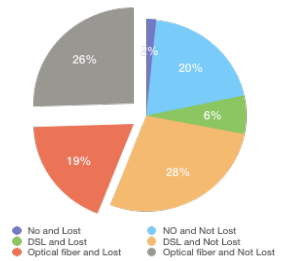


HAVING CHILDREN OR NOT	
No and Lost	1,549
NO and Not Lost	3,391
Yes and Lost	324
Yes and Not Lost	1,779

图 6 分析用户流失比例（按照有多条线路）

结论：是否有多条线路整体对用户流失影响不明显。

按照多条线路互联网服务提供商（DSL，Fiber optic，No）分析用户流失比例，如图 7 所示。

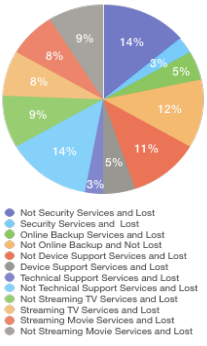


INTERNET SERVICE PROVIDER	
No and Lost	113
NO and Not Lost	1,408
DSL and Lost	457
DSL and Not Lost	1,958
Optical fiber and Lost	1,311
Optical fiber and Not Lost	1,796

图 7 分析用户流失比例（按照多条线路互联网服务提供商）

结论：光纤用户的流失占比较高。

根据互联网服务用户绑定情况分析用户流失比例，如图 8 所示。



BINDING OF INTERNET SERVICE USERS	
Not Security Services and Lost	1,461
Security Services and Lost	295
Online Backup Services and Lost	523
Not Online Backup and Not Lost	1,233
Not Device Support Services and Lost	1,211
Device Support Services and Lost	545
Technical Support Services and Lost	310
Not Technical Support Services and Lost	1,446
Not Streaming TV Services and Lost	942
Streaming TV Services and Lost	814
Streaming Movie Services and Lost	818
Not Streaming Movie Services and Lost	938

图 8 分析用户流失比例（根据互联网服务用户绑定情况）

工程技术与管理 · 第 04 卷 · 第 05 期 · 2020 年 05 月
DOI: https://doi.org/10.26549/gcjsygl.v4i5.4189

结论：由图可以看出，在网络安全服务、在线备份业务、设备保护业务、技术支持服务、网络电视和网络电影六个变量中，没有互联网服务的客户流失率值是相同的，都是相对较低。这可能是因为以上六个因素只有在客户使用互联网服务时才会影响客户的决策，这六个因素不会对不使用互联网服务的客户决定是否流失产生推论效应。

绑定了安全、备份、保护、技术支持服务的流失率较低；附加流媒体电视、电影服务的流失率占比较高。

根据付款方式分析用户流失比例，如图 9 所示。

PAYMENT METHOD	
Bank Transfer Payment and Lost	257
Bank Transfer Payment and Not Lost	1,285
Credit Card Payment and Lost	231
Credit Card Payment and Not Lost	1,289
Electronic Check Payment and Lost	1,070
Electronic Check Payment and Not Lost	1,293
Payment By Mail Check and Lost	315
Payment By Mail Check and Not Lost	1,303

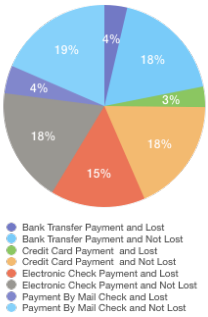


图 9 分析用户流失比例（根据付款方式）

结论：在四种支付方式中，使用 Electronic check 的用户流流失率最高，其他三种支付方式基本持平，因此可以推断电子账单在设计上影响用户体验。

根据消费额情况分析用户流失比例，如图 10 所示。

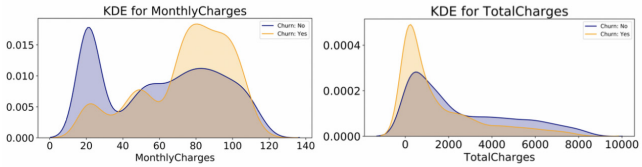


图 10 分析用户流失比例（根据消费额情况）

结论：月消费额大约在 70−110 之间用户流失率较高。

长期来看，用户总消费越高，流失率越低，符合一般经验。

3 将结果可视化

我们利用 Html 和 CSS 简单制作了一个网页，将上述所得结果呈现给用户，网页地址为 http://47.93.163.14。

4 结语

针对性给出增加用户黏性、预防流失的建议。

推荐老年用户采用数字网络，且签订 2 年期合同（可以各种辅助优惠等营销手段来提高 2 年期合同的签订率），若能开通相关网络服务可增加用户粘性，因此可增加这块业务的推广，同时考虑改善电子账单支付的用户体验。

电信业的竞争重点集中在对客户市场的争夺，这要求各大电信运营商将更多的精力投入到客户市场。做好客户的培育、巩固和回流工作，这三个方面的工作是相互促进、相互补充的。针对客户的回流工作，可采取以下措施以尽可能地降低客户的流失率。

4.1 开展个性化服务

现阶段企业服务水平的差异不是体现在大众化服务上而是体现在个性化服务上。目前电信消费群体对个性化消费的要求越来越高、电信企业如何适应消费群体定制化服务的要求，将特色服务作为企业新的竞争力和业务增长点，是电信企业迫切需要解决的问题。例如：针对年老、单身、无家属的用户推出特制服务，如人文套餐等，一可以增强用户之间的联系度，二可以提供个性化设计服务。

4.2 做好客户的开发和维持工作

良好的客户关系对于项目的成功有着不可低估的作用。及时掌握客户的通信需求，可以增进人与人之间的情感交流与思想沟通等，企业间的合作最终是人与人之间的合作，例如赠送半年或一年打折券。对于使用光纤和附加流媒体电影、电视服务的用户，重点在于提高网络使用体验、增值服务体验。

4.3 完善自身业务能力

电信运营商对现有的业务做好进一步的宣传工作。对客户需要而企业暂时不能开放提供的业务则要加大内部研发工作，不能让需求在等待中消失，更不能因能力不足而失去收入增长的机会。此外还要认真研究市场，做好业务的预测工作。

参考文献

[1] 许乃利 . 基于大数据技术的电信客户流失预测模型研究及应用 [J]. 信息通信技术 ,2018,v.12;No.63(02):68−73.

[2] 王观玉 , 郭勇 . 支持向量机在电信客户流失预测中的应用研究 [J]. 计算机仿真 ,2011(004):115−118,312.

[3] 王纯麟 , 何建敏 . 基于 AdaBoost 的电信客户流失预测模型 [J]. 价值工程 ,2007(02):112−115