Operations Research and Fuzziology 运筹与模糊学, 2022, 12(4), 1383-1391
Published Online November 2022 in Hans. http://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2022.124146

Hans汉斯

基于Cox模型的网路视频客户流失研究

杜前程，曾　进，张雨豪

贵州大学数学与统计学院，贵州　贵阳

收稿日期：2022年10月5日；录用日期：2022年11月2日；发布日期：2022年11月8日

摘　要

网络视频用户数量决定网络视频服务商收益，如何有效地降低客户流失率成为网络视频商家的关注重点。以和鲸社区公开的网络视频客户流失数据为研究样本，运用R语言软件对样本数据进行描述性统计，结合不同的解释变量对生存时间做对比描述分析。然后，对样本数据里的变量建立Cox模型，发现“未订阅电视、没有电影套餐、过去3个月账单平均值(15~30元)、因服务失败而呼叫中心的次数(0次)、过去3个月平均下载量(40 GB以上)、过去3个月平均上传量(3 GB以上)”的客户，流失可能性更低，生存时间更长。

关键词

Cox模型，网络视频，客户流失

Research on Network Video Customer Churn Based on Cox Model

Qiancheng Du, Jin Zeng, Yuhao Zhang

School of Mathematics and Statistics, Guizhou University, Guiyang Guizhou

Received: Oct. 5th, 2022; accepted: Nov. 2nd, 2022; published: Nov. 8th, 2022

Abstract

The number of online video users determines the revenue of online video service providers, and how to effectively reduce the customer churn rate has become the focus of online video businesses. Taking the online video customer churn data published by Hejing Community as the research sample, using R software to descriptive statistics of the sample data, and combining different explanatory variables to make a comparative description and analysis of the survival time. Then, build a Cox model for the variables in the sample data, and find that Customers with “no TV subscription, no movie package, average bill in the past 3 months (15~30 yuan), number of calls to the

文章引用: 杜前程, 曾进, 张雨豪. 基于 Cox 模型的网路视频客户流失研究[J]. 运筹与模糊学, 2022, 12(4): 1383-1391.
DOI: 10.12677/orf.2022.124146

center due to service failure (0), an average download volume of the past 3 months (above 40 GB) and an average upload volume of the past 3 months (above 3 GB)" are less likely to churn and survive longer.

Keywords

Cox Model, Network Video, Customer Churn

1. 引言

随着互联网技术的高速发展，截至 2021 年 12 月，我国网民规模达 10.32 亿，互联网普及率达 73.0% [1]。在今天，伴随着互联网科技的高速发展和进步，多屏互动技术也应运而生，它可以关联各种智能设备，包括平板电脑、智能手机、智能电视等，并可以即时完成信息共享。多屏互动技术[2]将媒介文本推发给不同的端口用户设备进行传播，或者将设备的显示屏投射到其他端口用户设备，能够通过无线互联互通的方法，让机器设备双方能够互相达到最优化地实现和共享资源。据统计，截至 2021 年 12 月，在网民中，网络视频、短视频用户使用率分别为 94.5%和 90.5%，用户数量分别达 9.75 亿和 9.34 亿[3]。网络视频市场前景非常可观，网络视频商家之间存在着激烈的竞争，拥有更多的用户也成为各网络视频服务商的首要目标。研究表明，获取新客户的成本是维护老客户成本的 5~6 倍[4]。所以，要获得更多的客户，不仅要考虑吸纳新客户，更需要维护现有客户。客户流失成为各企业关注的核心问题，如何有效地增强现有客户粘性的问题亟待解决。

客户流失问题的研究重点分为二个方向：客户流失预警与流失因素分析，预测将要流失的客源，并深入分析流失的因素，有针对性地制定挽留对策，以尽量地降低客户的流失。国内外研究者也对客户流失问题做过大量的研究，张宇等[5]结合邮政企业短信业务，建立 C5.0 决策树算法模型预测客户流失，该模型可以对短信产品客户流失的状况进行分析、保有月预警，具有较高的命中率和覆盖率，在一定程度上可以帮助企业尽可能减少客户流失。胡永培等[6]对银行优质客户的流失预警，先通过 AP 聚类算法实现属性筛选，将属性相似的归为一类，然后使用随机森林方法构建客户流失预警模型，预测零售优质客户未来 3 个月流失的风险，在评估结果中有较好的效果。夏国恩等[7]将改进的多层感知机应用在客户流失预测，对于在传统的客户流失预测数据预处理中，使用 one-hot 编码处理离散属性数据会使数据维度增加和数据过于稀疏的问题，提出了堆叠自编码器和实体嵌入两种方法对多层感知机进行改进，改进后的模型有效提高了预测的准确度，但是该模型中每个离散属性对应的嵌入层大小需要不断调试。Monica 等[8]构建基于深度神经网络的客户流失预测模型，应用于银行客户流失预测，取得较好的效果，但是神经网络模型的构建需要大量参数，黑盒操作，预测结果解释性不强。因此，从业务角度考虑，优先选择统计解释性强且预测效果好的算法进行客户流失预测，如基于统计学习理论的方法[9]。

国内外研究者对众多的客户流失现象开展了深入研究，但关于网络视频客户流失现象的问题，还尚未作出定量分析。本文通过运用生存分析中的 Cox 模型对网络视频客户应用生存问题做出预估，以追踪每个客户从接入到流失的完整流程，并预估客户流失时间以及流失的具体因素，为网视频服务商精准推广提供针对性的参考。

2. Cox 模型简介

2.1. 生存时间

生存时间[10]是指被观测的时间，按失效事件发生的日期为失访的最后一个被观察记录，一般用符号 t 表示，它是一个随机变量，取值大于 0，往往用生存函数、概率密度函数以及风险函数来描述它的分布特征情况。

生存函数(survival function)，常用 $S(t)$表示，是个体生存时间大于时间 t 的概率，记：

$$S(t)=P(T>t)=1-F(t) \tag{1}$$

其中，$F(t)$是指个体的生存时间 T 的分布函数，且有 $t=0$ 时，$S(t)=1$，$t=\infty$时，$S(t)=0$。

概率密度函数(probability density function)又叫作密度函数，该函数的图形为密度曲线，在任何时间段内死亡的比例和死亡出现的概率峰值均可从密度曲线找出，函数表达式为：

$$f(t)=\lim_{\Delta t\to 0}\frac{P(\text{个体在}(t,t+\Delta t)\text{内死亡})}{\Delta t} \tag{2}$$

危险率函数(hazard function)，又名为风险函数、瞬间死亡率、死亡强度、条件死亡率、危险率等，危险率函数是生存分析最基本的函数，即

$$h(t)=\lim_{\Delta t\to 0}\frac{P(t\le T\le t+\Delta t\,|\,T\ge t)}{\Delta t}=\frac{f(t)}{S(t)} \tag{3}$$

其中，$S(t)$是生存函数，$f(t)$是概率密度函数。

从(3)式，可以看出，$S(t)$和 $f(t)$的关系可以记为

$$S(t)=\exp\left\{-\int_0^t h(t)\mathrm{d}s\right\} \tag{4}$$

2.2. Cox 模型

Cox 模型[11]全名为 Cox 比例风险模型，是由英国统计学家 Cox 在一九七二年提出的。Cox 模型由于在表达形式上和参数模型很类似，但对各参数进行估计时又不依赖特定分布的假设，所以又称为半参数回归模型[12]。其基本形式如下：

$$h(t,T)=h_0(t)\exp\left\{X^{\mathrm{T}}\beta\right\} \tag{5}$$

式中，$h_0(t)$是一个同解释变量 X 无关的基准风险函数(baseline hazard function)，β 为回归系数，X 是预后变量，即协变量。基于 Cox 模型的假设，由于每个预后变量的危险函数在时间上与基准危险函数 h_0 成正比，从而不需要计算 h_0，应用起来非常简单。与此同时，相应的生存函数为：

$$S(t,T)=\left\{S_0(t)\right\}^{\exp\left(X^{\mathrm{T}}\beta\right)} \tag{6}$$

其中，$S_0(t)$为 t 时刻的基准生存函数

在时间因素和协变量的共同影响下，个体风险函数与基准风险指数之间的比率与时间无关，因为它们的比率并不随时间的变动而改变；而基准风险函数则仅跟时间相关，与解释变量无关。

Cox 模型对协变量参数 β 的估计一般采用似然函数法，而观测对象 i 在时刻 t_i 被观测到的似然函数表达式如下：

$$L_i(\beta)=\frac{h(t_i,X_i)}{\sum_{j:t_j\ge t_i}h(t_i,X_j)}=\frac{h_0(t_i)\theta_i}{\sum_{j:t_j\ge t_i}h_0(t_i)\theta_j}=\frac{\theta_i}{\sum_{j:t_j\ge t_i}\theta_j} \tag{7}$$

其中 $\theta_j = \exp\left(X_j^{\mathrm{T}}\beta\right)$，若观测对象之间相互独立，则 β 的偏似然函数为

$$L(\beta) = \prod_{i=1}^{n} L_i(\beta) = \prod_{i=1}^{n} \theta_i \left(\sum \theta_j\right)^{-1} \tag{8}$$

将式(8)写成对数形式，再对 β 求偏导后令其为 0，具体形式如下

$$\frac{\partial \ln L(\beta)}{\partial \beta} = 0 \tag{9}$$

求解式(9)便可得 β 的估计。

3. 基于 Cox 模型的网络视频客户流失分析

3.1. 数据来源与说明

本文使用和鲸社区公开的网络视频客户流失数据，共 72,274 条真实用户数据，经过数据预处理将资料不完整的样本删除，最后有 71,893 个资料完整的客户样本，其中流失客户数为 40,050 个，流失客户占比为 55.70%。根据研究需要以及数据的可获得性，选取了是否订阅电视、是否有电影套餐、过去 3 个月账单平均值、因服务失败而呼叫中心的次数、过去 3 个月平均下载量(GB)、过去 3 个月平均上传量(GB)、客户是否流失、服务年限 8 个变量进行分析。每个变量的定义、取值范围及类型如表 1 所示。

Table 1. Description of research data
表 1. 研究数据说明

变量类型	变量名	详细说明	取值范围	备注
因变量	服务年限	定量变量，某个客户已经使用该网络视频服务多长时间	0~12.8	单位为年，建模时和是否流失一起组合成因变量
	是否流失	定性变量	流失，未流失	流失占比 55.70%
解释变量	是否订阅电视	定性变量	订阅，未订阅	订阅占比 81.52%
	是否有电影套餐	定性变量	有电影套餐，没有电影套餐	没有电影套餐占比 33.41%
	过去 3 个月账单平均值	定量变量	0~406	单位为元，建模时，处理为定性变量：0~15 元、15~30 元、30 元以上三组，分别占比 37.61%，54.85%，7.54%
	因服务失败而呼叫中心的次数	定量变量	0~19	建模时，处理为定性变量：0 次、0 次以上两组，分别占比 83.49%、16.51%
	过去 3 个月平均下载量(GB)	定量变量	0~4415.2	建模时，处理为定性变量：0~20、20~40、40 以上三组，分别占比 41.82%、19.34%、38.84%
	过去 3 个月平均上传量(GB)	定量变量	0~453.3	建模时，处理为定性变量：0~1、1~3、3 以上三组，分别占比 34.64%、26.19%、39.18%

3.2. 描述性统计

运用 R 语言软件对样本数据进行描述性统计，结合不同的解释变量对生存时间做对比描述分析。首先从是否订阅电视开始，根据该变量的两个不同取值(订阅、未订阅)，将样本分成两组。每组分别估计生存函数，然后绘制在一张图中进行对比分析，如图 1 所示。从中可以观察到：订阅电视所对应的红线一直都在未订阅电视的蓝线下面，说明对任何时间点而言，订阅电视的生存概率都要低于未订阅电视。因此，未订阅电视的客户似乎粘性更高，流失的可能性更低，而订阅电视的客户流失的可能性更高。

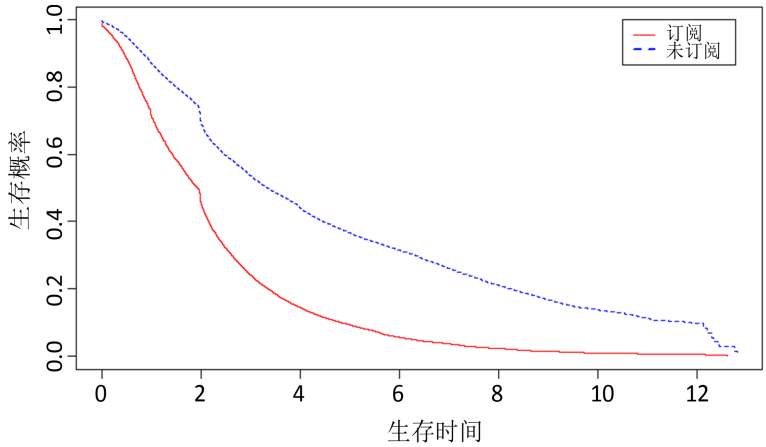

Figure 1. Survival curve of TV subscription or not
图 1. 是否订阅电视的生存曲线

其次，对另一个解释变量是否有电影套餐做类似的分析，结果如图 2 所示。从中可以看到一个清晰的规律：没有电影套餐对应的蓝线一直在有电影套餐的红线之上，说明对于任意一个时间点而言，没有电影套餐客户的生存概率都要明显一致地高于有电影套餐客户。因此，没有电影套餐客户似乎粘性更高，流失的可能性更低，而有电影套餐客户流失的可能性更高。

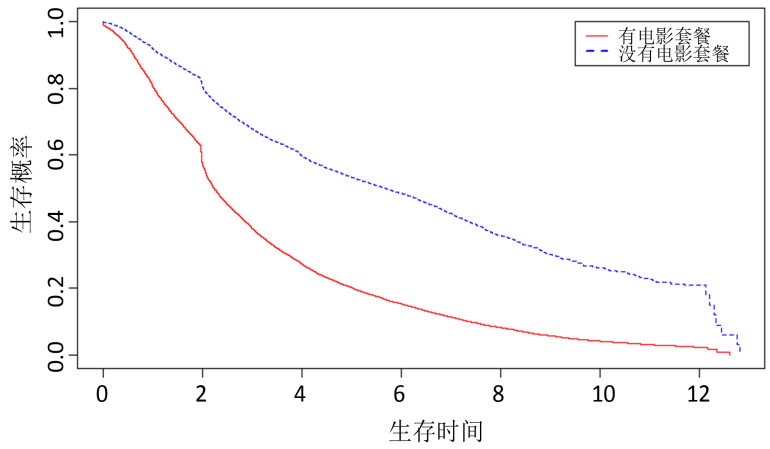

Figure 2. Survival curve of whether there is a movie package
图 2. 是否有电影套餐的生存曲线

再次，用同样的方法考察过去 3 个月账单平均值对生存时间的影响，结果如图 3 所示，从中可以观察到一个清晰的规律，15~30 元这一组对应的蓝线在其他两组(0~15 元)和(30 元以上)对应的线之上，说明

对于任意一个时间点而言，15~30 元这一组客户的生存概率要比另外两组高。因此，过去 3 个月账单平均值在 15 元~30 元的客户对网络视频服务商来说粘性更高，而另外两组(0~15 元)和(30 元以上)之间差别不大。

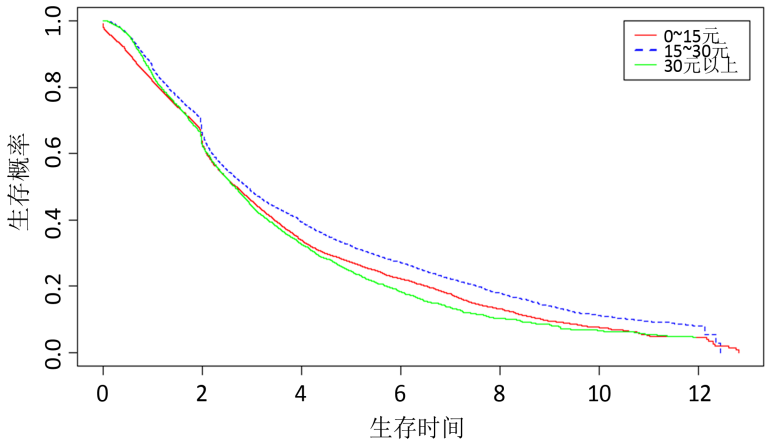

Figure 3. Survival curves of different groups of average bills in the past 3 months
图 3. 过去 3 个月账单平均值不同分组的生存曲线

然后，对变量因服务失败而呼叫中心的次数做类似的分析，结果如图 4 所示，从中可以看出，因服务失败而呼叫中心的次数(0 次以上)这一组对应的蓝线和因服务失败而呼叫中心的次数(0 次)这一组的红线之间的差别不大，表明这两组的生存概率几乎一致。因此，它们的客户粘性和流失的可能性相差无几。

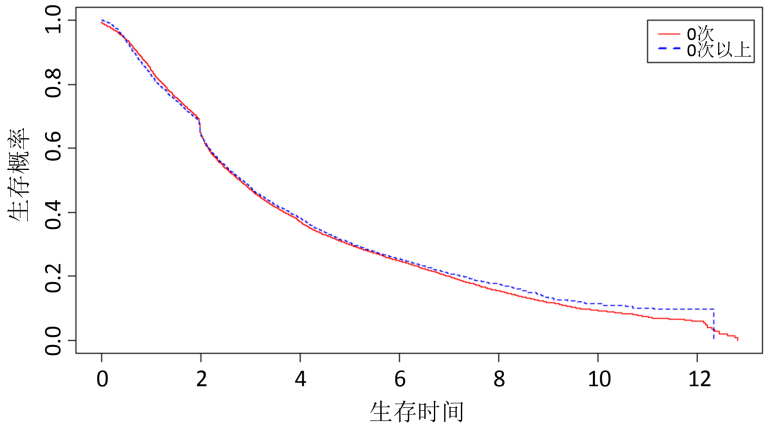

Figure 4. Survival curves of different groups of call centers due to service failures
图 4. 因服务失败而呼叫中心的次数不同分组的生存曲线

类似地，用同样的方法考察过去 3 个月平均下载量(GB)对生存时间的影响，结果如图 5 所示，从中可以看到一条明显的规律，40 GB 以上这一组对应的绿线一直在其他两组(0~20 GB)和(20~40 GB)对应的线之上，20~40 GB 这一组对应的蓝线一直都在 0~20 GB 这组对应的红线上面，说明对任何时间点而言，40 GB 以上这一组的生存概率最高，其次是 20~40 GB 这一组，最后是 0~20 GB 这一组。因此，40 GB 以上这一组客户的粘性最高，流失的可能性最低。而 0~20 GB 这组客户的粘性最低，流失的可能性最高。

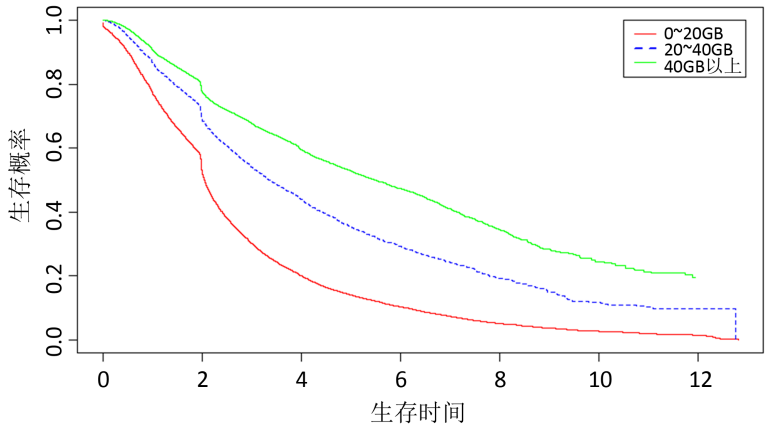

Figure 5. Survival curves of different groups of average downloads (GB) in the past 3 months
图 5. 过去 3 个月平均下载量(GB)不同分组的生存曲线

最后，对另一个解释变量过去 3 个月平均上传量(GB)做类似的分析，结果如图 6 所示，从中可以看到一条清晰的规律，3 GB 以上这一组对应的绿线一直在其他两组(0~1 GB)和(1~3 GB)对应的线之上，1~3 GB 这一组对应的蓝线一直都在 0~1 GB 这组对应的红线上面，说明对任何时间点而言，3 GB 以上这一组的生存概率最高，其次是 1~3 GB 这一组，最后是 0~1 GB 这一组。因此，3 GB 以上这一组客户的粘性最高，流失的可能性最低。而 0~1 GB 这组客户的粘性最低，流失的可能性最高。

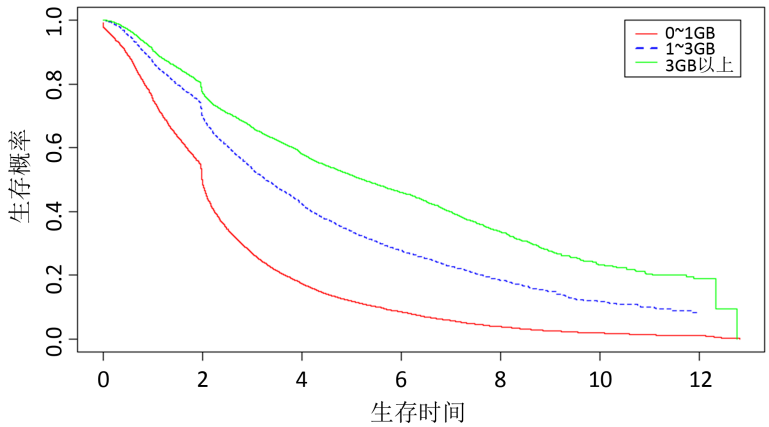

Figure 6. Survival curves of different groups of average uploads (GB) in the past 3 months
图 6. 过去 3 个月平均上传量(GB)不同分组的生存曲线

3.3. 模型估计及解释

运用 R 语言软件对所有解释变量建立 Cox 模型，结果如表 2 所示。模型整体检验 P-值小于 0.001，高度显著，说明所有解释变量中至少有一个是显著影响流失风险的。观察所有的解释变量相应的 P-值，发现除了过去 3 个月账单平均值(15~30 元)外其余都小于 0.001，因此大部分解释变量高度显著。根据表 2 可知，发现过去 3 个月账单平均值(30 元以上)、因服务失败而呼叫中心的次数(0 次以上)这两组对应的参数估计显著为正，说明这些变量组所代表的客户相比其照组的客户流失风险更大：过去 3 个月账单平均值(30 元以上)的流失风险是过去 3 个月账单平均值(0~15)元的 1.309 倍，因服务失败而呼叫中心的次数(0 次以上)的流失风险是因服务失败而呼叫中心的次数(0 次)的 1.089 倍；是否订阅电视、是否有电影套餐、过去 3 个月账单

Table 2. Cox model results
表 2. Cox 模型结果

变量	回归系数	exp(coef)	标准误	P-值	备注
是否订阅电视	−0.304	0.738	0.012	<0.001	无
是否有电影套餐	−0.641	0.527	0.013	<0.001	无
过去 3 个月账单平均值(15~30 元)	−0.027	0.973	0.011	0.0138	基本组：0~15 元
过去 3 个月账单平均值(30 元以上)	0.269	1.309	0.020	<0.001	
因服务失败而呼叫中心的次数(0 次以上)	0.085	1.089	0.013	<0.001	基本组：0 次
过去 3 个月平均下载量(20~40 GB)	−0.184	0.832	0.018	<0.001	基本组：0~20 GB
过去 3 个月平均下载量(40 GB 以上)	−0.447	0.640	0.022	<0.001	
过去 3 个月平均上传量(1~3 GB)	−0.413	0.662	0.016	<0.001	基本组：0~1 GB
过去 3 个月平均上传量(3 GB 以上)	−0.571	0.565	0.022	<0.001	
模型全局检验			<0.001		

平均值(15~30 元)、过去 3 个月平均下载量(20~40 GB)、过去 3 个月平均下载量(40 GB 以上)、过去 3 个月平均上传量(1~3 GB)、过去 3 个月平均上传量(3 GB 以上)、过去 9 个月的限制次数(0 次以上)这七组对应的参数估计显著为负，表明这些变量组所代表的客户相比对照组的客户流失风险更小：未订阅电视的流失风险是订阅电视的 0.738 倍，没有电影套餐的流失风险是有电影套餐的 0.527 倍，过去 3 个月账单平均值(15~30 元)的流失风险是过去 3 个月账单平均值(0~15 元)的 0.973 倍，过去 3 个月平均下载量(20~40 GB)和(40 GB 以上)的流失风险分别是过去 3 个月平均下载量(0~20 GB)的 0.832 倍、0.640 倍，过去 3 个月平均上传量(1~3 GB)和(3 GB 以上)的流失风险分别是过去 3 个月平均上传量(0~1 GB)的 0.662 倍、0.565 倍。

4. 结论

本文以和鲸社区公开的网络视频客户流失数据作为研究对象，先对其进行描述性分析，然后建立 Cox 模型，对客户流失的可能性进行预测，主要得到如下结论：

1) 结合不同的解释变量对生存时间做对比描述分析，发现未订阅电视客户的流失可能性比订阅电视客户的低，没有电影套餐客户的流失可能性比有电影套餐客户的低，过去 3 个月账单平均值在 15~30 元客户的流失可能性最低，因服务失败而呼叫中心的次数在 0 次以上客户的流失可能性比 0 次的低，过去 3 个月平均下载量在 40 GB 以上客户的流失可能性最低，过去 3 个月平均上传量在 3 GB 以上客户的流失可能性最低。

2) Cox 模型结果显示，在 5%的显著性水平上，过去 3 个月账单平均值(30 元以上)、因服务失败而呼叫中心的次数(0 次以上)这两个变量与网络视频客户流失正相关，是否订阅电视、是否有电影套餐、过去 3 个月账单平均值(15~30 元)、过去 3 个月平均下载量(20~40 GB)、过去 3 个月平均下载量(40 GB 以上)、过去 3 个月平均上传量(1~3 GB)、过去 3 个月平均上传量(3 GB 以上)这七个变量与客户流失负相关。

需要说明的是，各视频服务商家存在不同的发展条件与运营模式，其中分析出的导致用户流失的显著性原因不可能应用于任何视频服务商。尽管如此，我们所构建预测客户流失的 Cox 模型对视频服务商客户关系管理人员来说依然存在着参考价值。一方面，视频服务商家可以密切关注这些预测因素的变化，使其影响程度维持在一定范围，并朝着有利趋势发展，防止产生实质性转变而造成经济损失；另一方面，视频服务商可以应用 Cox 模型，根据当前已知的客户资料，给出相关协变量具体的数值，较精确的预测出视频服务商在未来某个时间节点上的客户流失状况，并提前采取最有效的对策方法。

参考文献

[1] CNNIC 发布第 49 次《中国互联网络发展状况统计报告》[J]. 新闻潮, 2022(2): 3.

[2] 徐滢, 杏运. 多屏互动专利技术综述[J]. 中国科技信息, 2021(19): 24-25.

[3] 武晓莉. 中国网民规模达 10.32 亿[N]. 中国消费者报 2022-03-03(003).

[4] Hadden, J., *et al*. (2005) Computer Assisted Customer Churn Management: State-of-the-Art and Future Trends. *Computers and Operations Research*, **34**, 2902-2917. https://doi.org/10.1016/j.cor.2005.11.007

[5] 张宇, 张之明. 一种基于 C5.0 决策树的客户流失预测模型研究[J]. 统计与信息论坛, 2015, 30(1): 89-94.

[6] 胡永培, 张琛. 基于 AP 聚类与随机森林的客户流失预测研究[J]. 计算机技术与发展, 2021, 31(2): 49-53.

[7] 夏国恩, 唐琪, 张显全. 改进的多层感知机在客户流失预测中的应用[J]. 计算机工程与应用, 2020, 56(14): 257-263.

[8] Hedge, S. and Mundada, M.R. (2019) Enhanced Deep Feed Forward Neural Network Model for the Customer Attrition Analysis in Banking Sector. *International Journal of Intelligent Systems and Applications* (*IJISA*), **11**, 10-19. https://doi.org/10.5815/ijisa.2019.07.02

[9] Ebrah, K. and Elnasir, S. (2019) Churn Prediction Using Machine Learning and Recommendations Plans for Telecoms. *Journal of Computer and Communications*, **7**, 33-53. https://doi.org/10.4236/jcc.2019.711003

[10] Walters, S.J. (2012) Analyzing Time to Event Outcomes with a Cox Regression Model. *Wiley Interdisciplinary Reviews*: *Computational Statistics*, **4**, 310-315. https://doi.org/10.1002/wics.1197

[11] Cox, D.R. (1972) Regression Models and Life Tables (with Discussion). *Journal of the Royal Statistical Society*: *Series B*, **34**, 187-220. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x

[12] 邓森文, 马溪骏. 基于 Cox 模型的移动通信行业中低端客户流失预测研究[J]. 合肥工业大学学报(自然科学版), 2010, 33(11): 1698-1701.