arXiv:2304.00575v1 [stat.ML] 2 Apr 2023

基于深度学习序列框架的零售业客户流失建模

Juan Pablo Equihua[1], Henrik Nordmark[1,3], Maged Ali[2] 和 Berthold Lausen[1,4*]

[1]埃塞克斯大学数学科学系，英国科尔切斯特。[2]埃塞克斯大学埃塞克斯商学院，英国科尔切斯特。[3]创新部门，Profusion Media Ltd.，保罗街，英国伦敦。[4]埃尔朗根-纽伦堡弗里德里希-亚历山大大学医学院医学信息学、生物统计学与流行病学研究所，瓦尔德街6号，德国埃尔朗根，邮编91054。

*通讯作者。电子邮箱：berthold.lausen@fau.de；其他作者：je18890@essex.ac.uk；henrikn@profusion.com；maaali@essex.ac.uk；

摘要

随着全球零售商加大力度，针对不同受众制定营销活动，准确预测哪些客户最有可能流失，对于营销团队提升业务利润至关重要。本研究提出一种深度生存分析框架，用于预测在非契约型场景中，哪些客户可能停止向零售公司购买商品。通过利用由循环神经网络学习的生存模型参数，我们能够仅依据个体客户行为，建立客户购买行为的个体级生存模型，从而避免训练机器学习模型时通常需要耗费大量时间进行的特征工程。

关键词： 客户流失、深度学习、生存分析、循环神经网络

引言

寻找创新方法以减轻客户流失、提高客户留存率，长期以来一直是金融业、技术行业、银行业、保险业等众多领域企业积极开展的工作。尤其是在零售领域，Reichheld (1990);Van den Poel 和 Larivi`ere(2004年) 等人开展的研究表明，获取新客户的成本通常是留住现有客户的 5 到 12 倍，尽管这一比例因公司而异。营销人员认识到，投入资源开展企业投入大量精力识别哪些客户可能会流失，以便为客户制定更具竞争力的营销策略。Reichheld（1990）研究表明，客户留存率仅提高5%，就能使银行业利润增加85%、保险经纪业利润增加50%、汽车行业利润增加30%。遗憾的是，导致客户流失的深层原因可能因企业和行业而异，尽管营销人员已指出，糟糕的客户服务、缺乏吸引力的价值主张、低质量的沟通方式以及品牌互动不足，是导致主动客户流失的四个主要原因。

公司在零售等非契约型场景中识别未来可能流失的客户，常用的方法是运用统计学和机器学习方法，预测哪些客户可能在特定时间段内停止与公司进行交易。在这类场景中，客户无需订阅，也可以不通知公司便改变购买习惯。随后，可针对流失概率较高的个体开展一项或多项客户挽留活动；这类活动通常提供专门设计的产品促销，以激励客户购买并持续与品牌互动，正如Borah、Prakhya 和 Sharma（2019年）所述。尽管预测客户流失这一概念相对直观，但设计此类系统时面临的实际问题可能颇具挑战，原因有多种。首先，在零售这类非合约场景中，客户可能随时改变购买习惯；通常，客户距离下一次购买的时间越长，其再次购买的概率就越低。其次，使用多变量统计方法和机器学习技术，需要为每位客户提取由领域专家通常提出的人工构造的特征。这些特征的质量、数量和类型会直接影响用于识别流失客户的分类方法的最终表现。Bengio、Courville 和 Vincent（2013年）。此外，由于客户的行为在不同领域和不同公司之间各不相同，要找到一组表现较好且能在不同公司间良好泛化的特征可能并不容易，因此，每逢新的情境，都可能需要耗时地寻找此类特征。

为解决这些问题，一些研究探索了采用人工智能、人工神经网络和表征学习等方法，在无需人工干预的情况下获取客户表征；其主要目标是避免公司在设计机器学习模型时必须进行的、耗时的特征工程步骤，并基于这样一种假设：正如Davenport、Guha、Grewal 和 Bressgott (2019年) 所指出的，人工智能将在未来十年改变营销策略和客户行为。例如，Spanoudes 和 Nguyen(2017年) 使用抽象特征向量和四层神经网络，预测哪些客户可能会在按月界定的时间范围内流失。然而，这种方法以及类似方法（如 Coussement 和 Van den Poel (2008年)；S.-Y. Hung、Yen 和 Wang(2006年)；Tamaddoni、Sepehri、Teimourpour 和 Choobdar(2010年) 提出的方法）并未考虑到在观察期内，关注事件可能尚未在所有个体身上发生这一情况，而文献中提出的基于生存分析的技术则考虑到了这一点，例如 Kaplan-Meier 估计。这种方法也被用于预测客户流失，因为其设计本身就适用于对事件在一段时间内发生的概率进行建模，无需像 Gul 等人(2020年)；Jamal 和 Bucklin (2006年)；Wong (2011年) 那样，为表征客户而获取大量特征。

为了有效预测客户流失并设计有针对性的营销活动，重要的是设计更有效的方法，同时考虑两方面因素：每位客户固有的购买行为，以及由不确定性所导致的删失效应，即无法区分哪些客户已经结束了与企业的关系，哪些客户只是在两次交易之间暂时处于停顿期。传统的预测近期可能流失客户的技术，通常只会考虑这两项要求中的一项。

本研究提出了一种新方法，用于对非契约型场景（如零售业）中的客户流失进行建模，主要有三个目标。首先，超越采用从客户行为信息中提取的人工设计特征的传统机器学习方法和基于生存分析的方法；具体做法是以基于深度学习的方法取代特征工程阶段，并利用循环神经网络估计模型参数。其次，通过排除人口统计数据等客户信息，并完全依据客户购买发生所需的时间来定义客户行为，从而减少营销模型中固有的算法偏差。最后，通过获取

每位客户到达时间的个体层面分布。

本文其余部分安排如下：第二节回顾利用深度学习方法和生存分析进行客户流失预测的现有文献。第三节概述本文采用的方法以及分析所需的数据预处理。第四节介绍基于一家英国大型零售商提供的真实交易数据集的实验结果。最后，第五节给出结论和未来工作方向。

文献综述

生存分析与事件发生时间建模

生存分析（SA）是统计学的一个分支，专注于对未来存续时间内的事件发生时间数据进行建模，即估计事件在未来某一特定时间之后发生的概率。回归和基于树的模型等机器学习（ML）方法假定所有事件在观察时均已发生，与之不同，生存分析假定，对于某些个体，事件在评估时可能尚未发生，但如果延长观察期，事件可能在未来发生，这种现象称为右删失。生存分析估计截至时间 t 结局事件尚未发生的概率，并通过生存函数 $S(t)$ 考虑数据中存在的删失，该函数定义为

$$S(t) = P(T > t) = 1 - F(t),$$

其中， T 是根据事件随时间的分布定义的随机变量，而 $F(t)$是事件时间的累积分布，通常根据一组个体属性 X （又称协变量、预测变量或特征）进行建模，即 $S(t \mid X) = P(T > t \mid X) = 1 - F(t \mid X)$。对于生存函数 $S(t)$，风险比或风险函数 $\gamma(t)$定义了时间 t的事件发生率；若事件截至时间 t 尚未发生，则风险函数表示为

$$\gamma(t) = \lim_{\triangle t \to 0} \frac{P(T < t + \triangle t \mid T \geqslant t)}{\triangle t} = \frac{S'(t)}{S(t)},$$

其中， $S'(t) = -f(t)$，而 $f(t)$是事件时间分布的概率密度函数。

生存分析方法可分为三大类：1）参数方法，它假设生存时间服从特定分布，并为模型协变量指定函数形式。2）半参数方法，它规定协变量与生存函数之间存在函数关系，但不对风险函数的具体形式作出限定。3）非参数方法，它既不对生存函数作任何假设，也不对协变量的分布作出假设。在文献提出的各种技术中，行业应用中使用的两种常见方法是 Cox 比例风险模型（CPH）Cox(1972年）和 Kaplan 和 Meier (1958年）估计量，因为它们具有灵活性，且易于实施。CPH 半参数形式允许将多个协变量的分布与基线风险进行线性组合，从而获得个体在时间 t 的风险概率估计。而 Kaplan-Meier 非参数结构则使公司和研究人员能够以简便且可扩展的方式，获得稳健的基线以及对个体随时间变化的风险进行可靠预测。Kaplan-Meier 估计量 $S(t)$ 的个体生存函数定义为

$$\hat{S}(t) = \prod_{k:t_k<t} \left(1 - \frac{d_k}{n_k}\right),$$

其中， d_k 是在时间 t_k 经历事件的个体数量， n_k 是在时间 t_k处于风险中的个体总数。

生存模型因其简单灵活，能够将多个协变量纳入风险函数估计，已被世界各地的多家公司和营销人员广泛用于预测客户流失。 Van den Poel 和 Larivi`ere(2004年) 探讨了在欧洲金融服务领域使用比例风险模型来建模客户流失。Wong (2011年) 使用 Cox回归 来识别影响加拿大某电信公司客户留存的人口统计和时间协变量。Jamal 和 Bucklin(2006年) 将客户服务、支付和回收系统中的时间依赖协变量与 Weibull 风险函数结合，用于识别南美洲某卫星电视服务的流失客户。 Mavri 和 Ioannou

(2008年) 研究了希腊银行业客户转换行为的潜在预测因素。

然而，比例风险模型并不直接对生存概率建模，而是对个体在时间 t 的风险函数建模，即 $\gamma(t \mid X) = \gamma_0(t)exp(X^T\beta)$。该函数表示在某个时间区间开始时事件尚未发生的前提下，个体在该时间区间内经历关注事件的概率。它由仅取决于时间 γ_0 的基线风险，以及根据个体协变量得到的时间无关函数 $X^T\beta$ 共同确定，其中 β 是与每个个体协变量相关联的 n 维权重向量。已知风险函数后，即可利用累积风险函数 $\Gamma(t) = \int_0^t \gamma(s)ds$ 得到生存函数，即 $\hat{S}(t) = \exp(-\Gamma(t)$。

通常，比例风险模型分两步进行拟合 Kvamme、Borgan 和 Scheel (2019年)。首先，通过最大化 Cox部分似然，拟合仅取决于个体协变量的模型参数部分 $exp(X^T\beta$，因为它不依赖于基线风险函数 γ_0；随后，根据上一步得到的参数估计结果，对非参数基线进行估计。

尽管生存分析技术已在多个应用场景中证明能够准确预测客户流失，但这些方法也有其局限性。首先，这些技术通常假设关注事件只能发生一次，并且只要观察个体的时间足够长，该事件就必然会发生；然而，在客户行为建模中，个体在其一生中可能会多次购买，因此这一假设并不成立 Mavri 和 Ioannou(2008年）；Spanoudes 和 Nguyen (2017年）；Tamaddoni 等人 (2010年），也可能在一段时间后不再购买。应对这些问题的一种务实方法是，在客户再次购买后立即将其生存概率重置为 1，并引入一个治愈因子，表示个体不再进行任何购买的概率 Amico 和 Van Keile- gom (2018年）。

其次，将生存模型应用于购买行为时，性能评估可能并不直接，因为生存分析技术估计的是事件随时间发生的概率，而评估传统客户购买行为则是对“客户最终会购买”与“客户再也不会购买”这一二元分类问题进行建模。这会导致某些方案在单个时间点上可能是最优的，但整体上并非如此。

最后，比例风险模型仍要求对所有个体使用恒定风险函数；对于非契约型零售业中的购买行为建模而言，这一假设并不现实，因为客户随时可能改变其购买模式。不同研究尝试通过多种方式克服这一难题：在原始 Cox 模型之上采用混合模型Nagpal 等人 (2019年），在对抗框架中训练比例风险模型 Chap- fuwa 等人 (2018年），或使用随时间变化的风险函数Fisher 和 Lin (1999年），或将生存模型与基于网络的架构相结合。例如，Nagpal、Li 和 Dubrawski (2020年）提出了一种名为 深度生存机器 的全参数化机制，无需采用强假设风险函数，即可学习协变量的非线性表示。而 Ren 等人(2018年）提出了一种深度循环生存模型，用于预测事件发生的可能性；该模型不对生存函数作任何特定的分布假设，同时考虑了数据中的删失情况。

生存分析中的《神经网络》

人工神经网络

人工神经网络（人工神经网络）是广受欢迎的机器学习模型，能够利用架构设计中相互关联的节点和加权连接，学习数据中复杂的非线性模式；这种架构设计主要受到人脑结构的启发。已有研究证明，神经网络模型在多个领域的表现优于若干机器学习技术，例如逻辑回归和基于树的模型Meng-hani (2021年)，例如自然语言处理（自然语言处理）Camacho-Colladosand Pilehvar (2017年）、计算机视觉Rawat andWang (2017年）和机器翻译 Sutskever,Vinyals, andLe (2014年）。此外，Kvamme 等人Kvammeet al. (2019年)和 Sharma 与 PanigrahiSharma andPanigrahi (2013年)也广泛探索了人工神经网络在生存分析和客户流失预测领域的应用，研究的任务中，观测数据可能存在删失，

并且可以从输入数据中提取协变量。

从数学角度看，神经网络（$NN : \mathbb{R}^n \rightarrow \mathbb{R}^m$）具有 n维输入和m维输出，可以被视为 $X \in \mathbb{R}^n$ 特征（节点）及其对应权重 $W \in \mathbb{R}^{n\times m}$和偏置项 $b \in \mathbb{R}$（连接）的线性组合或函数，随后接一个非线性激活函数，例如Sigmoid函数或双曲正切函数，即 $NN(X) = \varphi(b + W^T X)$，其中 $\varphi: \mathbb{R}^m \rightarrow \mathbb{R}^m$ 表示为网络选定的激活函数。在实践中，深度神经网络架构的输入层和中间层包含多个节点，使其能够在训练过程中学习从输入数据到输出数据的复杂非线性映射；该训练过程通过使用由Rumelhart 和 McClelland （1987）提出的反向传播算法或其变体，最小化网络输出与目标值之间的损失函数来完成。为简化表示，我们记 NN_X 为根据训练数据学习得到的神经网络 X。

虽然神经网络是强大的机器学习模型，但也有其局限性。首先，由于其结构通常包含多个中间层，模型可能含有数千甚至数百万个需要在训练过程中优化的参数，因此，为避免过拟合并提供可靠的预测，这些技术在训练时需要大量输入数据。其次，神经网络通常只提供目标变量的点估计，无法捕捉预测中的不确定性；当目标是估计概率分布（如本研究所做的那样）时，这并不理想。解决这些问题的一种常见方法是采用概率方法，并假设目标变量 Y 服从分布 $Y \sim Q(\theta)$，其未知参数向量为 θ；此时，神经网络的输出仅是对参数分布的估计，而非对目标变量的估计，即 $NN_X(x) = \varphi(b + W^T x) = \hat{\theta}$，目标变量的点估计则可通过计算该分布关于输入数据 $\hat{}$ 的期望得到，即 $\hat{y} = E[Q(\theta \mid x)]$。自然，这一过程不仅可以应用于输出层，也可以应用于部分或全部中间层，从而形成贝叶斯神经网络领域。这类网络通过最小化估计的后验分布与先验分布之间的Kullback-Leibler 散度 $Q(\theta)$，如 Lampinen 和 Vehtari (2001年 所述。

循环神经网络

循环神经网络（RNN）是传统网络架构的扩展，广泛应用于序列建模任务，例如 Yi、Nasukawa、Bunescu 和 Niblack 提出的文本分类 (2003年）、语言建模，以及 Sutskever 等人提出的机器翻译Sutskever 等人 (2014年）。由于其内部结构，RNN 可以将时变随机变量的信息编码到隐藏状态中 X_t，并利用该隐藏状态 h_t 定义时变目标变量的概率分布 Y_t，即 $P(Y_t \mid h_t)$。在 RNN 中，隐藏状态 h_t 随时间演变，利用 X_t 在每个时间步的信息，并通过函数 g将其与前一隐藏状态结合，即在每个时间步 $h_t = g(h_{t-1}, X_t)$，如图 1所示。

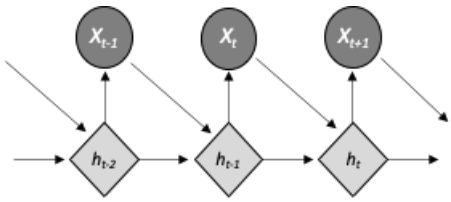

图 1 循环神经网络隐藏状态结构的图形表示。

然而，编码函数的选择g 并非易事，因为它需要能够捕捉序列数据中的长期依赖关系，同时避免在训练过程中发生过拟合。常用的选择是所谓的 g “记忆单元”函数，例如由Hochreiter 和 Schmidhuber (1997年) 提出的长短期记忆（LSTM），以及由 (GRU)Chung, G¨ul¸cehre、Cho 和 Bengio (2014年) 提出的门控循环单元。这些单元采用基于 Sigmoid 函数的内部门控机制，在循环训练过程中存储和遗忘信息。Bennis、Mouysset 和 Serrurier(2020年) 强调，与其他类型的模型相比，循环神经网络（RNN）在预测客户流失方面具有优势，尤其适用于从数据中提取的协变量会随时间变化的情形。

近期研究提出了将循环神经网络与生存分析相结合的方法，其表现优于传统的比例风险模型及其变体。在这些方法中，通常使用循环网络的输出来完全参数化生存模型，以预测未来事件的经验分布，并利用时间依赖协变量。Giunchiglia、Nemchenko 和 van der Schaar(2018年) 为医疗领域提出了 RNN-SURV。该方法获取患者一段时间内的特征，在每个时间步，模型都会以个性化的方式计算每位患者的风险评分和生存函数。Martinsson (2017年) 提出了 WTTE-RNN（威布尔事件发生时间循环神经网络），其主要做法是使用循环网络估计威布尔分布的参数，然后利用估计得到的威布尔分布预测机械设备维护领域中发动机的失效时间。Chen、Keng 和 Moreno(2018年) 提出了 MAT-RNN（多变量到达时间循环神经网络），将生存分析框架的预测扩展到多次到达的情形，例如在需求预测中预测购买行为。Ben-nis 等人(2020年) 提出了一种循环架构，用于对威布尔混合分布的参数进行建模，以开展事件发生时间分析。

非对称损失函数

通常，神经网络的训练过程是通过反向传播调整网络的权重和偏置，以找到能使模型预测值$\hat{y}$ 与真实观测值 y 之间误差最小的权重。此误差根据损失函数 $L(\hat{y}, y)$: $\mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$来衡量，该函数反映预测值与目标值之间的差异。在分类或回归任务中，目标变量完全已知，并由一组真实标签 y给出，因此模型会预测$\hat{y} = NN_X(x)$，模型的误差也可以直接计算。常用的损失函数包括用于回归的均方误差（$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$）和平均绝对误差（$MAE = \frac{1}{N}\sum_{i=1}^{N}|\ y_i - \hat{y}_i\ |$），以及用于分类的二元或分类交叉熵（$Entropy = -\sum_{i=1}^{N} y_i \cdot log(\hat{y}_i)$）。然而，这些损失函数没有考虑数据中存在删失标签的情况；只有当模型训练过程中对真实值的高估或低估不会造成显著影响时，它们才能提供可靠的估计，而在估计事件发生时间分布的参数时，情况并非如此。

过去几十年来，人们广泛研究了指数形状的事件发生时间分布参数估计，因为等待时间问题、事件之间的时间间隔等许多寿命问题通常都呈指数形状。Srivastava 和 Tanna (2007年）；Varian(1975年）；Zellner(1986年）等学者指出，与采用 MSE 和 MAE 等二次型损失进行参数估计相比，使用非对称损失函数的表现更好。这是因为其内在结构同时考虑了拟合优度与待估分布的吻合程度，以及各项估计的精确度。此外，Nelder 和 Wedderburn(1972年）提出了用于建模非对称分布的统计框架。

基于启发式思路提出的常见非对称损失函数包括 LINEX 损失函数Varian (1975年），该函数在零点一侧呈指数增长，另一侧呈线性增长。类似地，平衡损失函数（BLF）Zell-ner (1986年）将给定估计量与目标分布及其未知参数之间的距离结合起来。平衡损失函数通常用于贝叶斯统计领域，因为它可以纳入参数的先验知识，并以先验分布的形式加以表示。

方法论

本节介绍客户交易数据的数学表示，以及用于估计个体随时间进行购买的可能性的模型架构。我们的目标是训练一个神经网络，利用此前观测到的事件时间和删失事件时间的信息，估计指数分布到达时间的参数。随后，可使用该神经网络预测下一个事件时间，从而估计客户层面的生存分布。

数据表示

令 D_K 表示投资组合中 K 客户的所有购买记录集合。在非合约

场景下，购买可能发生在任意日期和时间，因此，将客户$k \in K$进行第 $i-th$ 次购买的日期记为 d_k, $_i$，如图3所示。通常，所有客户的跨国数据会整理到一个包含三列的单一数据集中，分别为“客户编号”“订单号”和“购买日期”，如表1所示。该数据集按客户编号分组，并按日期升序排列，这意味着，对于固定客户 k，所有 $i_1 \leqslant i_2$均满足d_k, $_{i_1} \leqslant d_{k,i_2}$。尽管第4节所述的一些事件发生时间方法会使用这种形式的数据来处理客户流失预测问题，但本研究还需要进行额外转换：定义随机变量 T_k，将其作为连续客户交易之间的时间差，即，对于$i \geqslant 2$，有$t_{k,i} = d_{k,i} - d_{k,i-1}$。假设这个新的随机变量 T_k 在客户层面服从指数分布，即 $T_k \sim Exp(\lambda_k)$，如图2所示。

为处理购买事件的右删失，我们考虑客户上次购买后至观察日期之间经过的时间，并将这一时间间隔定义为$t_{k,\ n_k+1} = analysis\ date - d_{k,\ n_k}$，作为所有客户的删失时间。

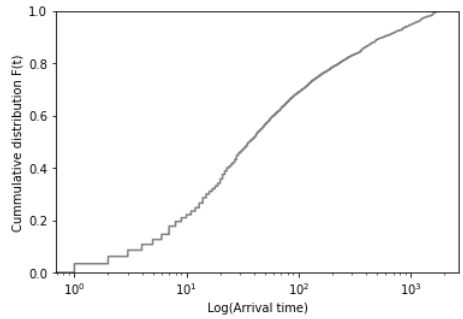

图2 客户连续购买之间的时间分布 $t_{k,\ \cdot}$（对数刻度）。

表1 原始交易数据

客户编号	订单号	购买日期
1	0	$d_{1,0}$
1	1	$d_{1,1}$
2	0	$d_{2,0}$
...	...	...
k	i	$d_{k,i}$
k	i+1	$d_{k,i+1}$
...	...	...

表 2 客户序贯交易

客户	到达时间序列	δ_k-序列
1	$[t_{1,1}, t_{1,2}, ..., t_{1,n_1}, t_{1,n_1+1}]$	[1,1, ..., 1, 0]
2	$[t_{2,1}, t_{2,2}, ..., t_{2,n_2}, t_{2,n_2+1}]$	[1,1,..., 1, 0]
...	...	...
k	$[t_{k,1}, t_{k,2}, ..., t_{k,n_k}, t_{k,n_k+1}]$	[1,1,..., 1, 0]

为了在序列框架中对 $k-th$ 客户的事件间隔时间 t_k 建模，我们将每位客户的所有到达时间t_k, $_j$ 压缩为一个新的序列向量 $\mathbf{t}$ $_k = [t_{k,1}, t_{k,2}, \ldots, t_{k,n_k}]$，每位客户对应一个 $k \in K$，其中每个到达时间$t_{k,j}$ 都假设服从指数分布。正如预期，每个序列的长度各不相同，取决于客户在其生命周期内于公司进行的购买次数，因此$length(\mathbf{t}_k) = n_k$，其中 n_k 是客户 k的购买总次数。接着，为了处理右删失事件时间（即由于客户可能在分析时点之后才进行下一次购买，因此仅被部分观察到的事件时间），我们将每位客户最后一次被观察到的时间与分析时点之间的间隔t_{k,n_k+1} 拼接到序列末尾，并为序列中的每个事件时间创建一个二元标识符$\delta_{k,j}$；事件被完整观察到时取值为$\delta_{k,j} = 1$，对于删失或部分观察到的时间则取值为$\delta_{k,j} = 0$。按此构造，每位客户 k 的序列 δ_k 在序列末尾只包含一个删失事件。训练数据 T_X 定义为所有独立 $\mathbf{t}_k$ 向量的并集，即 $T_X = \bigcup_{k \in K}(\mathbf{t}_k \cup t_{k,n_k+1})$，如表 2 所示，并附有相应的$\delta_{k,j}$ 标识符。实际应用中，这些 $\mathbf{t}_k$ 向量在训练和服务阶段都可以填充至任意长度。

训练过程

训练目标是开发一种基于机器学习的神经网络，假设每位客戶 $k \in K$ 的到达时间 t_k 服从指数分布，其参数为特定值 λ_k。为此，我们采用循环神经网络，并接入一个由单个输出单元构成的多层感知机，通过 ˆ Sigmoid 激活函数估计 $\lambda_k = NN_{T_X}(\mathbf{t}_k)$，随后利用该估计值对到达时间的指数密度函数 g 进行参数化

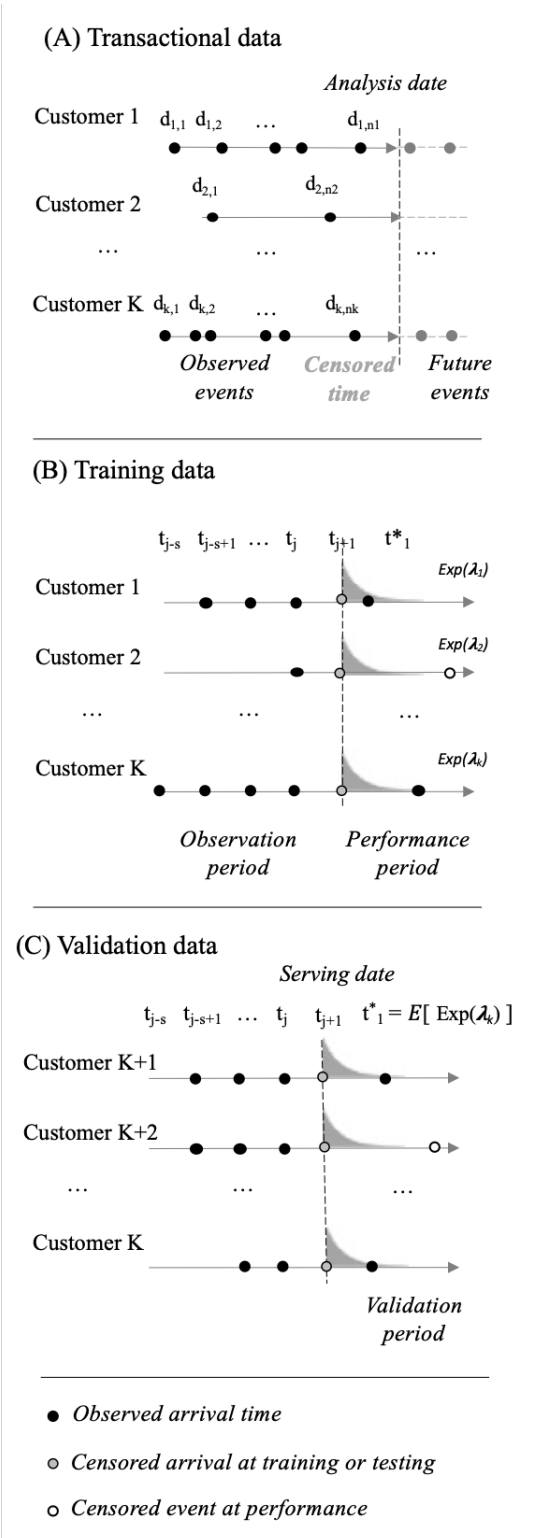

图 3 获得指数分布$t_{k,\ j} = d_{k,\ j} - d_{k,\ j-1}$所需的数据转换。每个客户序列均观测至时间 j，以预测 $t_{k,\ j+1}$。

对于每位客户。在每个 j-th 时间步，模型的输入数据是序列 $\mathbf{t}_{k,\ j} = [t_{k,j-s}, t_{k,j-s+1}, \ldots, t_{k,j}]$ 目标为 $t_{k,\ j+1}$，其中s 是序列填充参数，可根据不同应用任意设定，通常小于 $max(n_k)$，并在模型训练期间进行微调。图 4展示了所提出的最终模型架构。模型参数可以通过两种不同方式估计。首先，最小化加权非对称损失函数；该函数利用为每个观测创建的标识符 δ_k 来考虑删失事件，其中 $\delta_k = 1$ 对应完全观测到的事件时间，$\delta_k = 0$ 对应删失时间，同时对删失观测中较大的 $\hat{t}$ 预测值进行惩罚：

$$Loss = \left(\frac{1}{K_{\delta=1}} \sum_{k=1}^{K} \sum_{j=1}^{n_k} \omega_{t_k} (\hat{t_{k,j}} - t_{k,j})^2 \cdot 1_{\{\delta_{k,j}=1\}} \right.$$

$$\left. + \frac{1}{K_{\delta=0}} \sum_{k=1}^{K} \sum_{j=1}^{n_k} (1 - \omega_{t_k}) (\hat{t_{k,j}} - E[t])^2 \cdot 1_{\{\delta_{k,j}=0\}} \right)$$

其中，加权因子 $\omega_{t_k} = P(\delta_k = 1 \mid T_k = t)$ 表示训练数据中客户 k 在时间 t_k 的完全观测事件的概率，$E[t]$ 是目标分布 T 的估计期望值。该值可通过在每个训练批次中，在存在删失事件的情况下利用最大对数似然估计指数分布的参数来获得，具体如下：

$$\hat{\lambda} = \frac{K_{\delta=1}}{K \cdot \bar{t}},$$

其中 t 是观测事件时间和删失事件时间的均值 Kalbfleisch 和 Prentice （2002年）。

除了前述的非对称损失函数（正如第二节所述的类似方法中通常采用的那样），模型还可以通过最大化针对删失数据的模型部分似然函数进行训练，该函数表示为

$$L = \prod_{k=1}^{K} \prod_{j=1}^{n_k} f(t_{k,j} \mid \hat{\lambda}_k)^{\delta_{k,j}} \cdot S(t_{k,j} \mid \hat{\lambda}_k)^{1-\delta_{k,j}}$$

通常会将其转换为训练中的最小化任务，即最小化删失事件的负对数似然，其表示为

$$LL = -\sum_{k=1}^{K}\sum_{j=1}^{n_k}\left(log(f(t_{k,j} \mid \hat{\lambda}_k))1_{\{\delta_{k,j}=1\}} + log(S(t_{k,j} \mid \hat{\lambda}_k))1_{\{\delta_{k,j}=0\}}\right)$$

虽然该模型可以采用两种不同的方法进行训练，即使用非对称损失函数或最大似然作为损失函数。对于第四节中介绍的实验，我们决定使用负对数似然作为损失函数来训练模型，因为该函数考虑到了误差分布并不对称。

最后，在服务阶段，一旦已知 λ_k ，就可以轻松估计客户 k 在时间 t的生存概率，即

$$\begin{aligned} S_k(t) &= 1 - P(T_k < t) \\ &= 1 - \int_0^t \lambda_k exp(-\lambda_k x)dx \\ &= exp(-\lambda_k t) \end{aligned}$$

然后，我们在客户层面定义 $S_k(t)$的估计量如下:

$$\hat{S}_k(t) = exp(-\hat{\lambda}_k t)$$

$$S_k(t) = P(T_k > t \mid \lambda_k)$$

此外，为了计算每位客户在特定服务日期的生存状态，我们在每位客户的当前最近购买时间处评估 ˆ $S_k(t)$，即从客户上次购买到服务日期所经过的时间。此外，还可以通过简单计算 ˆ$S_k(t_1)$− ˆ $S_k(t_2)$，来获得一段时间内未来延迟事件概率，该概率表示下一次客户购买将在时间区间 (t_1, t_2) 内发生的概率。训练和预测步骤的伪代码描述见算法 1。

性能指标

本节简要介绍用于评估存在删失的评估数据中模型的主要方法。如前所述

算法1循环模型伪代码

输入： 表1所示的客户交易数据，包括'客户编号' 和'购买日期'。

输出： 用于预测下一个客户事件时间的循环神经网络。

将数据拆分为训练集、测试集和验证数据集：

- 将客户集合拆分为两个互不重叠的训练集和验证集，分别包含客户数据的 80% 和 20%。

- 根据所选分析日期，将训练客户交易拆分为两个互不重叠的训练数据集和测试数据集。

数据处理：
1. 获取每位客户后续购买之间的时间间隔，作为 $t_{k,i} = d_{k,i} - d_{k,i-1}$。2. 获取每位客户的序列输入向量，作为 $T_X = \bigcup_{k\in K} \mathbf{t}k \cup t_{k,n'_k}$

3. 使用滑动窗口遍历每个客户序列，将序列填充至相同长度，并将目标事件时间设为序列中紧接着的下一次到达时间。

模型训练：
Rep直到收敛 **g敛：**
1. 使用随机权重和偏置初始化循环神经网络。

2. 对输入序列执行网络前向传播，以获得每位客户下一次到达时间的估计值。

3. 使用实际观测到的和经删失处理的到达时间，以及估计的下一次到达时间，计算模型损失。

4. Up日期权 g重和网络中的偏置。

模型预测：
1. 为验证数据集创建输入序列 $T_X = \bigcup_{k\in K} \mathbf{t}$ $_k \cup t_{k,\ n'_k}$2. 对循环模型进行前向传播，以获得下一事件时间的估计值。3. 从每位客户的事件时间后验分布中抽取估计的下一事件时间 $t_{k,}$ $_{j+1}$ ，以估计每位客户的 λ_k 。4. 通过获取 S_k $(t) =$ˆ$exp(-\lambda_k t)$，预测客户生存概率的估计值ˆ

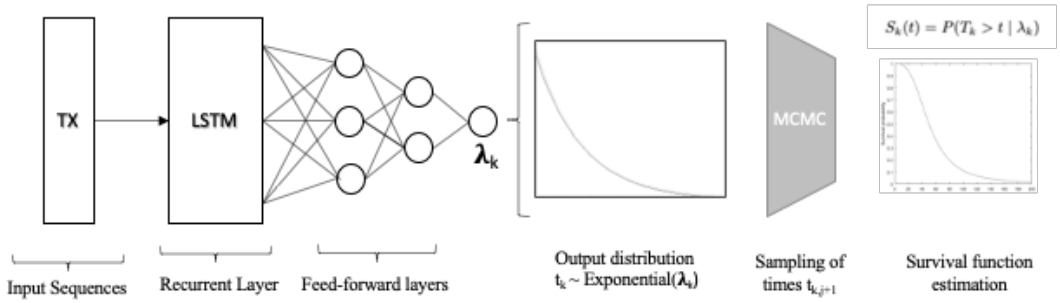

图 4 所提出的神经生存模型。交易序列输入经由长短期记忆单元（LSTM）和多层感知机处理，再通过大小为 1 的指数激活函数，为客户杠杆指数模型 $\mathbf{t}_k$ 参数化。在客户层面，生存模型 S_k 通过哈密顿采样从后验分布中抽取。

此前，在预测客户流失时考虑删失，会进一步增加整体模型评估的难度，因为我们无法充分区分已经流失、不会再购买的客户，与只是在两笔交易之间暂时停顿、但最终会回来的客户。

Brier 分数

Graf、Schmoor、Sauerbrei 和 Schumacher 提出的 Brier 分数（1999年）是生存分析中常用的评估指标，用于评估生存概率的准确率。它表示所有受试者的观测生存状态与预测生存概率之间的平均平方距离。在不存在删失的情况下，期望 Brier 分数可表示为

$$BS(t) = \frac{1}{K}\sum_{k=1}^{K}(1_{\{t_k^*>t\}} - S_k(t))^2,$$

其中 t_k^* 表示客户的首次到达时间，验证期内的客户 k 。然而，为了考虑生存模型中存在删失的情况，必须使用删失逆概率权重对评分进行调整。对于每个个体，考虑 $t'_k = min($ t_k, C_k) 以及 $\delta_k^* = 1_{\{t_k^* \leqslant C_k\}}$，其中 C_k 表示每个个体 k当前的观察时间。设 $G(t) = P(C > t)$ 表示时间 t的删失概率，通常通过 Kaplan-Meier 估计Graf等人（1999年）。在假设只要对个体进行足够长时间的观察，关注事件就必然发生的情况下，针对删失数据估计的时间依赖 Brier 分数定义如下：

$$BS(t) = \frac{1}{N}\sum_{k=1}^{N}\left(\frac{(0 - S_k(t))^2 \cdot 1_{\{t_k^* \leq t, \delta_k^*=1\}}}{G(t_k)} + \frac{(1 - S_k(t))^2 \cdot 1_{\{t_k^* > t\}}}{G(t)}\right)$$

最后，积分布赖尔分数（IBS）可对截至给定时间 t_{max} 的所有时点的模型性能进行总体估计。

$$IBS(t_{max}) = \frac{1}{t_{max}}\int_0^{t_{max}} BS(t)dt$$

一致性指数

Harrell一致性指数，也称为C指数或C统计量 Harrell、Califf、Pryor、Lee和Rosati (1982年)，是生存模型中最常用的性能指标之一，因为它的设计能够处理数据中的删失。与通过衡量预测误差来评估模型预测能力的指标（例如Brier 分数）不同，C指数通过比较可比个体对的预测评分与实际观测时间之间的相关性ˆ，来评估风险评分的区分能力 $S(t)$以及可比个体的实际观测时间 k_i 和

k_j，其中 $i \neq j$在不同时间 t_{k_i} 和 t_{k_j} 分别经历了该事件。C指数定义如下：

$$\text{C-Index} = P(S_{k_j}(t_{k_j}) > S_{k_i}(t_{k_i}) \mid t_{k_i} > t_{k_j})$$

C指数值越大，表明对哪些个体更容易经历所关注事件的预测越有信息量。

时变 ROC曲线下面积

受试者工作特征（ROC）是机器学习中一种成熟的技术，用于评估二分类器的分类能力，尤其是在目标变量的时间范围固定时（Fawcett，2006年）。对于任意分类器 Y： $X \rightarrow [0,1]$，其二元结果为$Y = \{0,1\}$，分类阈值为 c，ROC 曲线会比较分类器的敏感度（表示为 $sensitivity(c) = P(Y_1 > c \mid Y = 1)$，又称真正率（TPR））与 1 减去特异度，其中 $specificity(c) = P(Y_0 \leq c \mid Y = 0)$，又称假正率（FPR），并考察 Y 的所有可能取值；其中，Y_1 是正例的预测概率，Y_0 是在所定义的分类阈值 c 下负例的预测概率。因此，AUC（ROC曲线下面积）可定义为 ROC 曲线下方的总面积，并可解释为随机选取的一对观测值被 Y 正确分类的概率（Kamarudin、Cox 和 Kolamunnage-Dona，2017年），即：

$$AUC = P(\hat{Y}_1 > \hat{Y}_0)$$

AUC 是模型在所有可能分类阈值上的综合性能指标；AUC 较大表明分类器具有较强的预测能力，能够区分不同类别。

人们提出了不同的方法，将这一指标扩展到考虑可变的时间范围以及并非恒定的事件发生概率，例如 Heagerty和Zheng (2005年)；Kamarudin等人。 (2017年)；Lambert和Chevret(2016)。简单来说，当将ROC曲线扩展到时间依赖性结局时，敏感度和特异度都会相对于时间依赖性随机变量 T成为时间依赖性指标，其定义如下：“*t*时刻的累积病例”，即在 t之前经历事件的个体，$t_k^* \leq t$；以及 “*t*时刻的动态对照”，即在时间 t 之后经历事件的个体，$t_k^* > t$。因此，敏感度和特异度可以表示为时间t的函数，Heagerty和Zheng (2005年)如下：

$$sensitivity(c,t) = P(S_k(t) > c \mid t_k^* \leq t)$$

$$specificity(c,t) = P(S_k(t) \leq c \mid t_k^* > t)$$

因此，累积/动态ROC（C/D ROC）衡量模型在不同时间点对经历了事件的受试者进行分类的效果，而累积/动态AUC（C/D AUC）Heagerty、Lumley 和 Pepe（2000年）；H. Hung 和 Chiang（2010年）提供了一个单一的综合指标，用于表示C/D ROC曲线下的总面积，该指标代表在时间 t 时，已经经历事件的个体的非事件估计值高于尚未经历事件的个体的概率。估计的C/D AUC在时间 t 定义如下

$$AUC(t) = P(S_{k_i}(t_i) > S_{k_j}(t_j) \mid t_i \leq t < t_j)$$

如前所述，大多数生存分析技术都假设所有个体最终都会发生关注事件，因此，在 $t_2 < \infty$时将 $AUC(t_1,\ t_2)$作为性能指标，可能会产生评估偏差，而这种情况在实际应用中可能普遍存在。因此，如果关注事件未必会在所有个体身上发生，那么在生存分析应用中，Brier 分数和C指数等指标可能比时间依赖性曲线下面积更有用。

实验与结果

Profusion 是一家数据咨询公司，总部位于英国伦敦，为零售行业的众多企业提供数据与分析服务。

以及英国的金融行业。Profusion 使用不同的统计学和机器学习方法，在客户数据库中识别可能流失的客户。通常，每家公司每月进行一到两次流失估算；根据估算结果，可能流失的客户会被纳入营销活动，以主动鼓励他们使用公司的产品和服务。

本研究旨在提高为英国一家大型零售公司进行客户流失预测的分类能力。由于保密限制，该公司的名称不予披露，以下以 $company$ 指代。此外，我们还在一个合成数据集中评估模型性能，该数据集体现了零售业交易数据的主要特征。为比较我们的方法，我们建立了两个基线模型：一个仅使用交易数据的初始 CPH 模型，以及一个个体层面的 Kaplan-Meier 估计量。尽管 CPH 模型可以将年龄、性别和客户类型等客户层面特征纳入模型协变量，为了在仅使用交易数据评估模型性能时进行公平比较，本研究采用的 CPH 基线模型仅基于连续事件之间经过的时间。与前文所述流程类似，我们假设每位客户的删失事件时间由其最近一次购买日期与分析日期之间的间隔确定。

如前所述，比例风险模型并不直接对事件时间进行建模，而是建模个体在时间 t的风险函数，即个体在时间 t经历关注事件的概率；一旦已知风险函数，便可将个体的生存函数估计为 $S(t) = \exp(-\Gamma(t))$，其中 $\Gamma(t)$为累积风险函数。

此外，我们将模型性能与个体 Kaplan-Meier 基线进行比较；为此，我们通过考虑客户 k的所有观测事件时间和删失事件时间，估计每位客户的生存函数 S_k ，并针对其当前删失时间对其当前生存状态进行预测。

数据集 1：零售数据

第一个数据集包含$company$这家位于英国的大型零售商的交易数据，该零售商几乎拥有 200 家实体门店并提供在线订购服务。在 2016 年 3 月 29 日至 2021 年 4 月 20 日期间，该 $company$ 公司共完成了 1,780 万笔交易，涉及 280 万名客户。对于至少有 2 次购买的客户，观测事件之间的平均间隔为 80 天，即，考虑所有客户时，时间到事件指数模型的估计 λ 参数为 $\lambda = 0.012$。尽管我们掌握一些单个客户的人口统计信息，例如客户类型和所在地，但为了按照第三节所述对客户流失进行建模，我们在分析中仅纳入客户编号和购买日期，以获取事件之间的时间间隔。由于零售业属于非合约场景，该数据集显示，在 1400 天后，估计每月客户流失率为 14.7%，也就是说，14.7% 的客户将不再在该公司进行任何购买。图5展示了该数据集在 12 个不同月份滑动窗口下的累积对数生存概率。

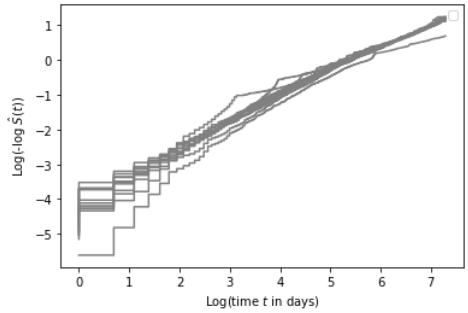

图 5 通过 Kaplan-Meier 曲线估计的零售数据集生存函数，展示了过去 30 天内进行过购买的客户的 12 个月度队列。

最后，在模型训练中，我们考虑了随机抽取的 100,000 位客户所进行的所有交易。

数据集 2：合成数据

按照 Bender、Augustin 和 Blettner (2005) 的方法，可以获得一个具有已知 λ_k 的、包含逼真多事件生存数据的合成数据集。为了获得足够大的数据集，我们模拟了一个包含 100,000 名客户的集合，其中每位客户的 λ_k 均从均值为 $\mu = 0.08$、标准差为 $\sigma = 0.02$的高斯分布中抽取。也就是说，该合成数据集中的预期客户回访时间为每 12.5 天一次。为了

避免 λ_k 出现负值，我们设定最低阈值为 0.01，因此， $\lambda_k = max(0.01, N(\mu, \sigma^2))$。随后，为了体现零售行业固有的非合约场景下客户不再继续购买的影响，我们在t_k的每次采样迭代中，为每位客户引入 15% 的停止概率，$j \sim exp(\lambda_k)$。对于该数据集，120 天后估算的月度客户流失率为 8.7%。图6展示了该数据集在 6 个不同月份滑动窗口下的累积对数生存概率。表3列出了两个实验数据集的主要汇总统计量，包括数据集规模、数据中事件的频率、训练期和绩效期，以及测试集划分规模。

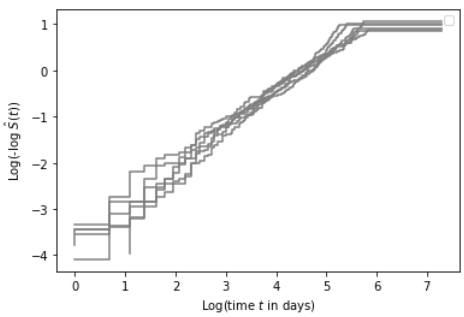

图 6 通过 Kaplan-Meier 曲线估计的合成数据集生存函数，涵盖过去 30 天内有过购买行为的客户的 12 个月度队列。

结果

如前所述，所分析的两个数据集中的购买间隔时间均服从指数分布。在所有情况下，客户进行下一次购买的概率都会随着时间推移而显著下降；距离上次购买 100 天后，仅发生了少量交易。表 3 列出了各数据集的汇总统计量。

所有数据集的数据预处理均按照第三节所述进行。对于每个数据集，我们将任意分析日期之前的交易数据压缩为客户层面的序列表示，从而获得一个形式与表格2类似的数据集，即对于每个数据集，我们创建如下训练数据： $T_X = \bigcup_K [t_{k,n_k-s}, t_{k,n_k-s+1}, ..., t_{k,n_k} \cup t_{k,n'_k}]$ 对于每个 $k \in K$，其中 s 是序列填充参数，用于在训练时将所有序列调整为相同长度。各数据集使用的这些填充参数见表3 。在序列数据建模中，通常应以这样的方式选择训练分析日期：绩效期完全不包含可用于服务的首个日期。在我们的实验中，所有数据集的分析日期都至少设在模型服务日期前 6 个月；此外，对于每个模型，我们指定一个验证集，其中的客户与训练时使用的客户完全不重叠。图3 展示了为每位客户创建向量$\mathbf{t}_k$ 所需的整体数据处理过程。

在预测时，我们按照第三节所述创建输入序列 $[t_{k,n_k-s}, t_{k,n_k-s+1}, ..., t_{k,n_k}, t_{k,n_k+1}]$，以估计 λ_k，针对性能评估数据集中的每位客户 $k \in K$；对于两项实验，该数据集均设定为训练集的补集。随后，可通过对由 λ_k 参数化的指数模型进行积分，估计任意时间 t 的生存概率 $S_k(T > t)$。图12 展示了分析时点购买频率处于四个不同细分群体中的客户个体生存曲线估计值。图 7展示了在相应验证数据集中，使用三种不同方法估计高、低购买频率客户生存函数的总体估计结果。此外，表 4列出了第三节所述指标在两个数据集上的性能结果。如图所示，循环模型在 Brier 分数方面优于两种基线方法；然而，C指数和时变 AUC 在 Cox 模型中显著更高。

正如预期，从图7 和图8 可以看出，循环神经网络模型能够较为有效地学习事件时间的生存分布，其结果接近于 $S(t)$ 的 Kaplan-Meier 估计。不过，由于事件时间分布假设事件时间服从指数分布，神经网络估计的整体生存函数 $S(t)$ 永远无法与 Kaplan-Meier 估计的生存函数完全吻合。由于

表3 两个实验数据集的主要汇总统计量。

	零售数据集	合成数据集
客户数量	2.8 M	100 K
训练集大小	10%	80%
验证集大小	5%	20%
观察期（MM/YYYY）	03/2016 - 05/2020	N/A
绩效期（MM/YYYY）	06/2020年 - 04/2021年	N/A
客户购买次数中位数	2	12
购买间隔中位数	32 天	12.5 天
填充序列的长度（s）	5	7

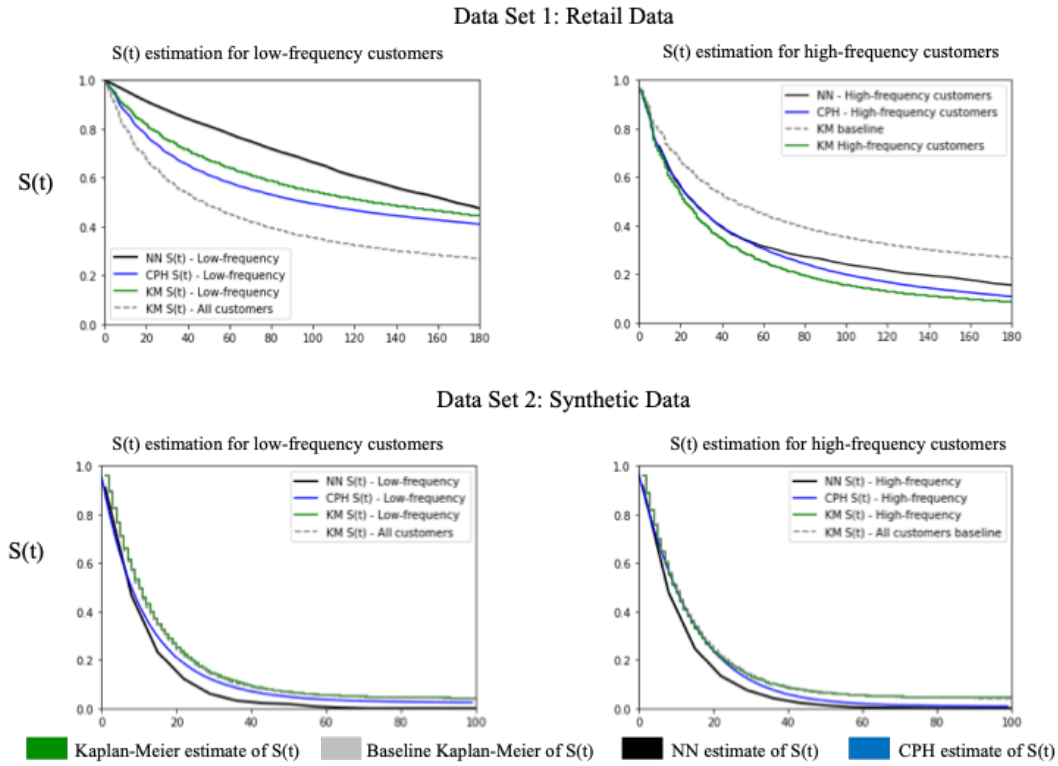

图7 对于购买频率较低和较高的客户，两个验证数据集中的平均客户的估计函数 $S(t)$。

合成数据集的设计方式是，事件频率和距最近一次事件的时间（距上次事件的时间）不会影响客户的流失率。因此，预计不同购买频率的客户，其估计生存曲线不会有太大差异，如合成数据集的图7所示。

此外，我们还使用零售数据集进行了分析，以研究我们采用神经网络估计生存概率的方法捕捉单个客户购买频率的变化。为此，我们分析根据每位客户的事件时间进行线性回归得到的斜率，并判断该斜率是较高（表明事件间隔时间延长、购买频率降低）、较低（表明事件间隔时间缩短、购买频率提高），还是保持不变（表明购买频率随序列没有变化）。图 10

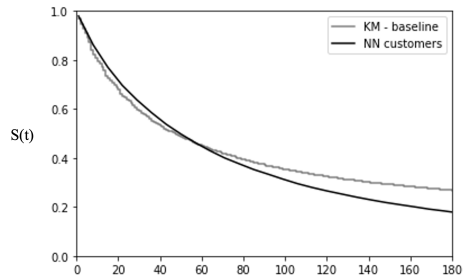

图 8 所提出的神经网络与基准 Kaplan-Meier 估计量在客户层面估计的生存函数$S(t)$的逐点平均值。

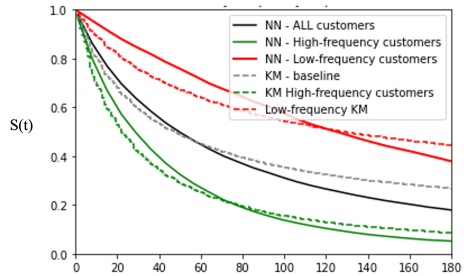

图 9 所提出的神经网络与基准 Kaplan-Meier 估计量针对高、低购买频率在客户层面估计的生存函数$S(t)$的逐点平均值。

展示了这三组客户的估计 λ 分布，图11 展示了 $S(t)$这三组客户的估计生存分布。

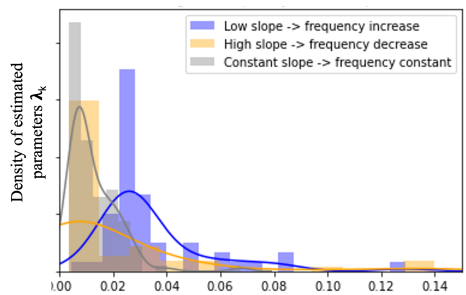

图 10 零售数据集中观察到购买频率变化的客户所对应的估计参数 λ_k 分布。

最后，我们通过将模型视为二元分类任务，评估模型在准确率、敏感度和特异度方面的表现。该任务的目标是利用客户的估计生存概率，预测事件是否

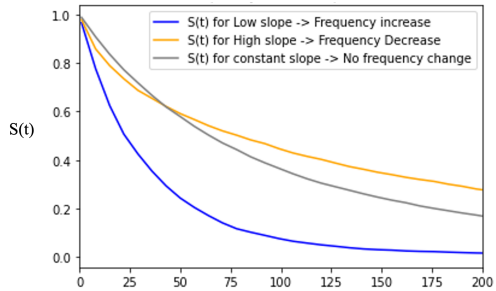

图 11 零售数据集中观察到购买频率变化的客户所对应的生存函数估计值的逐点平均值。

会在时间 t之前发生。为此，我们计算事件在时间t的中位生存概率，记为 $m(t)$，并使用一条分类规则：对于满足$S_k(t) < m(t$的观测，将其预测为正向事件，即事件可能较早发生在客户 k身上；否则预测为负向事件。表5 展示了此次评估的结果。

表4 在测试集和验证集划分中针对两个数据集进行实验所得的模型性能。C指数和时间依赖性曲线下面积越大，Brier 分数越小，表示模型性能越好。

	零售数据		合成数据	
	Test	验证	Test	验证
180天和100天时的 Brier 分数				
RNN	0.303	0.369	0.350	0.431
CPH	**0.070**	**0.197**	**0.021**	**0.186**
个体 Kaplan–Meier	0.384	0.437	0.378	0.292
C指数				
RNN	0.391	0.545	0.500	**0.498**
CPH	**0.580**	0.233	**0.675**	0.421
个体 Kaplan–Meier	0.555	**0.574**	0.495	0.465
时间依赖性曲线下面积				
RNN	0.348	0.458	0.487	**0.493**
CPH	0.259	0.307	0.364	0.482
个体 Kaplan–Meier	**0.591**	**0.590**	**0.494**	0.468

结论

世界各地的公司都想知道哪些客户可能会流失

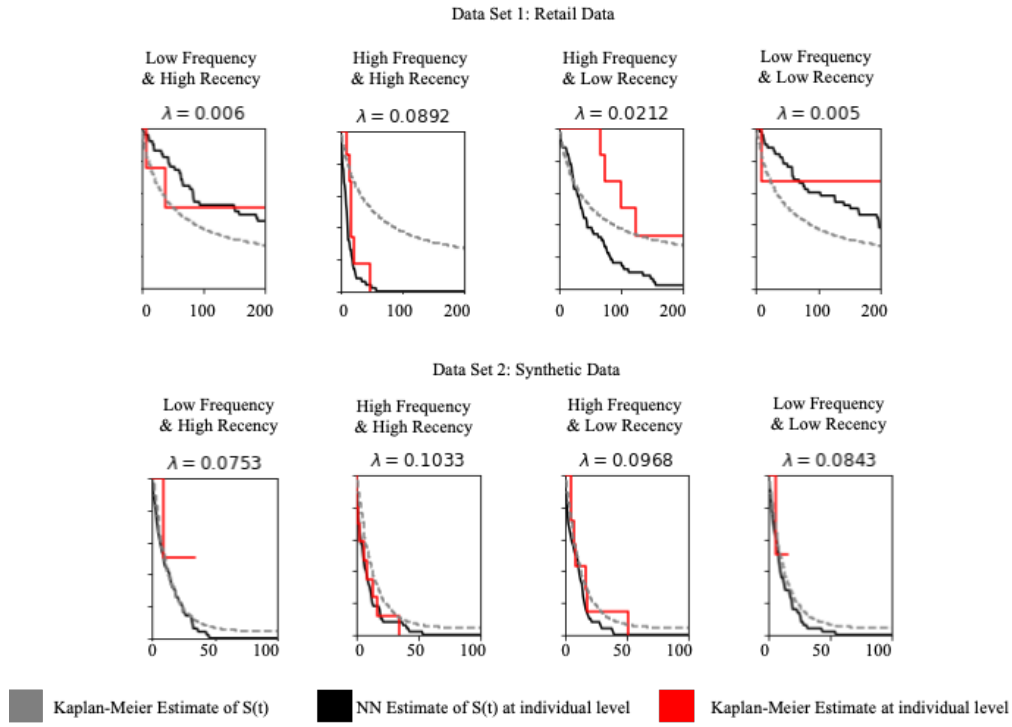

图 12 根据分析时点的购买频率和当前最近购买时间，对数据集中属于不同组的客户在未来 200 天和 100天内的客户层面估计生存函数。

表 5 实验中获得的模型性能
对两个数据集均进行，并考虑使用期望的
生存函数来预测事件是否会发生
在某个时间点之前或之后 t，$t = 180$ 天和 $t = 100$ 天，分别用于
零售数据和合成数据。作为二元
分类问题，准确率、
敏感度和特异度的值越高，模型预测能力就越强
。

	零售数据		合成数据	
	Test	验证	Test	验证
准确率				
RNN	**0.654**	0.525	0.560	0.210
CPH	0.485	0.462	0.366	**0.511**
个体 Kaplan–Meier	0.592	**0.646**	**0.712**	0.322
敏感度				
RNN	0.659	0.457	**0.883**	**0.833**
CPH	0.539	0.541	0.491	0.427
个体 Kaplan–Meier	**0.664**	**0.616**	0.780	0.769
特异度				
RNN	0.325	0.683	0.147	0.115
CPH	0.419	0.436	0.146	**0.623**
个体 Kaplan–Meier	**0.560**	**0.718**	**0.258**	0.227

以便主动采取客户留存措施，让他们持续关注品牌，并激励客户与产品互动。通过预测客户随时间进行下一次购买的概率，我们的模型能够为每位客户估算个体层面的生存函数，而非针对整个总体建立一个总体生存模型。

在购买时间建模中使用循环神经网络来预测客户流失，可以完全根据交易数据对客户的购买行为进行建模，而无需考虑年龄、性别和收入等客户层面特征。公司通常会利用这些特征估计客户与品牌互动并再次购买的可能性。此外，将购买时间建模为序列问题，并采用 LSTM 等循环网络架构，模型便能从历史交互中学习依赖关系，以最低限度的特征工程工作，或无需从输入数据中获取成本高昂的人工构造特征，达到或提升成熟生存分析技术的客户流失预测效果。

然而，我们的方法也有局限性。首先，假设每位客户的事件时间都服从指数分布，可能会导致低估时点 t 的剩余生存概率，尤其是对于最忠诚的客户细分群体而言。在这些群体中，即使观察期很长且没有购买，下一次购买的概率仍应保持较高。其次，该方法需要每位客户有大量购买记录，才能提供可靠的预测；对于购买历史较短的客户，模型的准确性可能较低。实验结果显示，模型在合成数据集上的表现更好，因为其中的购买频率远高于零售数据中的购买频率。因此，该方法更适用于客户购买频率较高的零售公司。尽管如此，对于事件时间较少的客户，我们的方法仍能在个体层面估计客户流失状态和生存概率；其他方法（如针对个体的 Kaplan-Meier 方法）无法做到这一点，而比例风险模型（CPH）等方法的估计则不够准确。

未来的研究可以探索该方法的泛化能力，例如将来自不同信号的多元时间序列作为模型中 LSTM 层的输入数据，以及压缩购买的季节性信息，或纳入有关所购商品的上下文信息；而这些信息在零售业中通常很容易获取。

声明

资助。本研究通过 Innovate UK、埃塞克斯大学（注册编号：Z699129X）与 Profusion Media LTD（在英格兰注册的公司，编号为 6947442）参与的知识转移合作计划获得支持。

利益冲突。作者无任何需要披露的相关经济或非经济利益。

代码可用性。本研究代码的知识产权归 ProfusionMedia 有限公司所有，不对外公开。

作者贡献。 J.E 实施了算法，进行了数据分析并撰写初稿；B.L. 核实了结果并监督了这项工作。所有作者均修订了论文，并阅读和同意手稿的已发表版本。

参考文献

Amico, M., & Van Keilegom, I.（2018年）。生存分析中的治愈模型。《统计学及其应用年度评论》，*5*(1)，311-342。取自 https://doi.org/10.1146/annurev-statistics-031017-100101

10.1146/annurev-statistics-031017 -100101

Bender, R., Augustin, T., Blettner, M.（2005）。用于模拟 Cox 比例风险模型的生存时间生成方法。医学统计学，*24*(11)，1713-1723。取自 https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.2059

https://doi.org/10.1002/sim.2059

Bengio, Y.、Courville, A.、Vincent, P.（2013年08月）。表征学习：综述与新视角。《*IEEE*模式分析与机器智能汇刊》，*35*，1798-1828。

10.1109/TPAMI.2013年.50

Bennis, A.、Mouysset, S.、Serrurier, M.（2020年2月）。使用神经网络进行事件发生时间分析中的条件混合威布尔分布估计（含右删失数据）。取自 https://hal.archives-ouvertes.fr/hal-02483979

Borah, S.、Prakhya, S.、Sharma, A.（2019年02月）。利用服务补救策略降低新兴市场中的客户流失。《营销科学学会期刊》，*48*。

10.1007/s11747-019-00634-0

Camacho-Collados, J. 与 Pilehvar, M.T.（2017）。论文本预处理在神经网络架构中的作用：关于文本分类和情感分析的评估研究。*CoRR*，*abs/1707.01780*。取自 http://arxiv.org/abs/1707.01780

Chapfuwa, P.、Tao, C.、Li, C.、Page, C.、Goldstein, B.、Carin, L.、Henao, R.（2018）。对抗性事件发生时间建模。机器学习研究论文集，*80*，735-744。Chen, T.、Keng, B.、Moreno, J.（2018年11月）。利用循环神经网络处理多元到达时间，以实现个性化需求预测。*2018*年 *IEEE* 国际数据挖掘会议研讨会（*ICDMW*）（第810-819页）。IEEE计算机学会。取自 https://doi.ieeecomputersociety.org/10.1109/ICDMW.2018.00121

Chung, J.、G¨ul¸cehre, C¸.、Cho, K.、Bengio, Y.（2014年）。门控循环神经网络在序列建模中的实证评估。*CoRR*，*abs/1412.3555*。取自 http://arxiv.org/abs/1412.3555

Coussement, K. 与 Van den Poel, D.（2008年）。订阅服务中的客户流失预测：在比较两种参数选择技术的同时应用支持向量机。*Expert Systems with Applications*，*34*(1)，313-327。取自 https://www.sciencedirect.com/science/article/pii/S0957417406002806

Cox, D.R.（1972年）。回归模型与生命表。*Journalofthe RoyalSta-tistical Society. Series B(Methodologi-cal)*，*34*(2)，187–220。取自 http://www.jstor.org/stable/2985181

Davenport, T.、Guha, A.、Grewal, D.、Bressgott, T.（2019年10月）。人工智能将如何改变营销的未来。《营销科学学会期刊》，*48*，1-19。

10.1007/s11747-019-00696-0

Fawcett, T.（2006年）。ROC分析简介。*PatternRecognitionLet-ters*，*27*(8)，861-874。获取自

https://www.sciencedirect.com/science/article/pii/S016786550500303X

Fisher, L.D., & Lin, D.Y.（1999年）。生存分析中的时变协变量与Cox比例风险回归模型。《公共卫生年度评论》，*20*(1)，145-157。取自 https://doi.org/10.1146/annurev. publhealth.20.1.145

Giunchiglia, E., Nemchenko, A., van der Schaar, M.（2018年）。RNN-SURV：用于生存分析的深度循环模型。人工神经网络与机器学习——*ICANN 2018*（第23–32页）。沙姆：施普林格国际出版。

Graf, E., Schmoor, C., Sauerbrei, W., Schumacher, M.（1999年）。生存数据预后分类方案的评估与比较。《医学统计学》，*18*(17-18)，2529—2545。取自 https://doi.org/10.1002/(sici)1097-0258(19990915/30)18:17/18;2529::aid-sim274;3.0.co;2-5

Gul, N.、Faiz, N.、Brawn, D.、Kulakowski, R.、Khan, Z.、Lausen, B.（2020年）。最优生存树集成。arXiv。取自 https://arxiv.org/abs/2005.09043

Harrell, F.、Califf, R.、Pryor, D.、Lee, K.、Rosati, R.（1982年）。评估医学检测的产出。《美国医学会杂志》，*247(18)*，2543-6。

Heagerty, P.J.、Lumley, T.、Pepe, M.S.（2000）。针对删失生存数据和诊断标志物的时变ROC曲线。《生物统计学》，*56*(2)，337-344。取自 https://onlinelibrary.wiley.com/doi/abs/10.1111/j.0006-341X.2000.00337.x

Heagerty, P.J., & Zheng, Y.（2005）。生存模型预测准确率和 ROC 曲线。《生物统计学》，*61*(1), 92-105。检索自 https://onlinelibrary.wiley.com/doi/abs/10.1111/j.0006-341X.2005.030814.x

Hochreiter, S., & Schmidhuber, J.（1997, 12）。长短期记忆。神经计算，*9*，1735-80。

10.1162/neco.1997.9.8.1735

Hung, H., & Chiang, C.-T.（2010年）。基于生存数据的时间依赖型 AUC 模型估计方法。《加拿大统计学评论》，*38*(1)，8–26。取自 http://www.jstor.org/stable/27805213

Hung, S.-Y., Yen, D.C., Wang, H.-Y.（2006年）。将数据挖掘应用于电信客户流失管理。《专家系统及其应用》，*31*(3)，515-524。取自 https://www.sciencedirect.com/science/article/pii/S0957417405002654

Jamal, Z., & Bucklin, R.E.（2006年）。改进客户流失的诊断与预测：一种异质风险建模方法。《互动营销杂志》，*20*(3)，16-29。取自 https://www.sciencedirect.com/science/article/pii/S1094996806700524

Kalbfleisch, J. 与 Prentice, R.（2002年，01）。《失效时间数据的统计分析》第二版，第 8 节。载于（第 247–277 页）。10.1002/ 9781118032985.ch8

Kamarudin, A.N.、Cox, T.、Kolamunnage-Dona, R.（2017年）。医学研究中的时间依赖性 ROC 曲线分析：当前方法与应用。《*BMC Medical Research Methodology*》，*17*（1），53。取自 https://doi.org/10.1186/s12874-017-0332- 6

Kaplan, E.L., & Meier, P.（1958年）。基于不完整观测的非参数估计。《美国统计学会杂志》，*53*(282)，457–481。检索自 http://www.jstor.org/stable/2281868

Kvamme, H., Borgan, Ø.,Scheel, I.（2019年）。使用神经网络和Cox回归进行事件时间预测。《机器学习研究杂志》，*20*，129:1-129:30。

Lambert,J., & Chevret, S.（2016）。基于累积/动态时变ROC曲线的生存模型判别能力汇总指标。《医学研究中的统计方法》，*25*(5), 2088-2102。获取自 https://doi.org/10.1177/0962280213515571

10.1177/0962280213515571

Lampinen, J., & Vehtari, A.（2001年）。神经网络的贝叶斯方法——综述与案例研究。《神经网络》，*14*(3), 257-274。获取自 https://www.sciencedirect.com/science/article/pii/S0893608000000988

Martinsson, E.（2017年）。*WTTE-RNN*：威布尔事件时间循环神经网络（博士学位论文或硕士论文）（博士学位论文，查尔姆斯理工大学，瑞典哥德堡）。获取自 http://publications.lib.chalmers.se/records/fulltext/253611/253611.pdf

Mavri, M., & Ioannou, G.（2008年，02）。运用生存分析研究希腊银行业服务中的客户转换行为。《管理金融》，*34*，186-197。

10.1108/03074350810848063

Menghani, G.（2021年）。高效深度学习：关于如何让深度学习模型更小、更快、更好的综述（第 abs/2106.08962 卷）。检索自 https://arxiv.org/abs/2106.08962

Nagpal, C., Li, X., Dubrawski, A.（2020年）。深度生存机器：针对含竞争风险的删失数据进行全参数生存回归与表征学习。*CoRR*, *abs/2003.01176*。检索自 https://arxiv.org/abs/2003.01176 Nagpal, C., Sangave, R., Chahar, A., Shah, P., Dubrawski, A., Raj, B.（2019年）。用于生存分析的非线性半参数模型。*CoRR*, *abs/1905.05865*。检索自 http://arxiv.org/abs/1905.05865 Nelder, J.A., & Wedderburn, R.W.M.（1972 年）。广义线性模型。《皇家统计学会期刊》。*A*辑（综合），*135*（3），370–384。检索于2022-04-30，网址：http://www.jstor.org/stable/2344614 Rawat, W., & Wang, Z.（2017年，06）。用于图像分类的深度卷积神经网络：全面综述。《神经计算》，*29*，1-98。

10.1162/NECO a_00990

Reichheld, F. (1990). 零客户流失——质量适用于服务。《哈佛商业评论》，9月至10月，105-111。摘自 https://ci.nii.ac.jp/naid/10010571997/en/ Ren, K., Qin, J., Zheng, L., Yang, Z., Zhang, W., Qiu, L., Yu, Y. (2018年). 深度循环生存分析。CoRR，abs/1809.02403。摘自 http://arxiv.org/abs/1809.02403 Rumelhart, D.E., & McClelland, J.L. (1987). 通过误差传播学习内部表征。载于《并行分布式处理：认知的微观结构探索：基础》（第318-362页）。

Sharma, A., & Panigrahi, P.K. (2013年). 一种基于神经网络的方法，用于预测蜂窝网络服务中的客户流失。*CoRR*，*abs/1309.3945*。取自 http://arxiv.org/abs/1309.3945

Spanoudes, P., & Nguyen, T. (2017年). 客户流失预测中的深度学习：基于抽象的公司无关特征向量进行无监督特征学习。*CoRR*，*abs/1703.03869*。取自 http://arxiv.org/abs/1703.03869

Srivastava, R., & Tanna, V. (2007年). 一般熵损失函数下指数寿命模型尺度参数的双阶段收缩估计量。统计学通讯：理论与方法，*36*(2)，283-295。取自 https://doi.org/10.1080/03610920600974104

10.1080/03610920600974104

Sutskever, I., Vinyals, O., Le, Q.V.（2014）。使用神经网络进行序列到序列学习。*CoRR*，*abs/1409.3215*。取自 http://arxiv.org/abs/1409.3215

Tamaddoni, A., Sepehri, M., Teimourpour, B., Choobdar, S.（2010年，12月）。非契约型环境中的客户流失建模：电信服务提供商案例。《战略营销杂志》，*18*，587-598。

Van den Poel, D., & Larivi`ere, B.（2004年）。运用比例风险模型分析金融服务中的客户流失。《欧洲运筹学杂志》，*157*（1），196-217。取自 https://www.sciencedirect.com/science/article/pii/S0377221703000699

Varian, H.（1975年）。一种用于房地产评估的贝叶斯方法。《贝叶斯计量经济学与统计学研究》，为……而作 *Leonard J. Savage*，195-208。

Wong, K.K.（2011年）。使用Cox回归对无线通信行业的客户流失时间进行建模。《营销定位、衡量与分析期刊》，*19*，37-43。

Yi, J., Nasukawa, T., Bunescu, R., Niblack, W. (2003年)。情感分析器：使用自然语言处理技术提取有关给定主题的情感。第三届 *IEEE* 国际数据挖掘会议(第427-434页)。10.1109/ICDM.2003.1250949

Zellner, A. (1986年)。使用非对称损失函数进行贝叶斯估计和预测。美国统计学会杂志，*81*(394)，446–451。取自 http://www.jstor.org/stable/2289234