基于客户-客服沟通文本信息的客户流失研究

王菲菲[1,2]，刘雯珺[3]，朱立奥[2]，吕晓玲[1,2*]

1.中国人民大学应用统计研究中心，北京 100872

2.中国人民大学统计学院，北京 100872

3.北京大学数学科学学院，北京 100871

摘要

客户流失问题对企业盈利能力产生重大影响，客户关系管理一直是公司的首要任务。因此，识别导致客户流失的原因并提出相应的运营管理策略，从而尽可能挽留客户，意义重大。服务行业普遍以一对一的形式为客户提供匹配客服的长期服务，客服人员对所服务客户具有深刻的了解，这些能为企业提前预判客户是否流失提供参考。随着信息技术的发展，数据记录手段的提升，客户-客服通过企业 APP 交流沟通的文字对话被详细记录，这些丰富的数据为研究客户流失原因提供了重要资源。以某互联网素质类在线教培公司的客户-客服对话数据为基础，通过文本分析方法挖掘对话中蕴含的各类主题以及情感状态，并进一步探索对话内容对客户流失的影响程度。结果显示，客户-客服沟通信息能够明显影响客户未来的续费表现，并能明显提升预测准确率。因此建议企业管理者注重客服沟通质量，站在消费者的视角，更好的挖掘消费者的消费意愿，从而与客户维持良好的业务关系保留客户。

关键词：客户关系管理；客户流失；客服服务；续费预测；文本分析

中图分类号：C931，C939

Research on Customer Churn Detainment based on Communication Text Information between Customers and Customer Services

Feifei Wang[1,2], Wenjun Liu[3], Li'ao Zhu[2], Xiaoling Lu[1,2]

1. Center for Applied Statistics, Renmin University of China, Beijing 100872, China

2. School of Statistics, Renmin University of China, Beijing 100872, China

3. School of Mathematical Sciences, Peking University, Beijing, 100871, China

* 通讯作者：吕晓玲（1977.8-），女、吉林人、管理学博士，E-mail：xiaolinglu@ruc.edu.cn.

本文受到国家自然科学基金项目(72001205, 72171229)、教育部人文社会科学重点研究基地重大项目(22JJD110001)、全国统计科学研究项目(2022LD06)的支持。

Abstract

The impact of customer churn on profitability makes customer relationship management become the primary goal. Therefore, it is of great importance to identify the churn status of customers and also develop management strategies and operations to retain customers. The service industries generally provide customers with long-term customer services in the form of one-to-one. Therefore, the customer service staff usually have a deep understanding of the customers they have served. As the development of technology, data recording methods are largely improved. It makes the textual communication information between customers and customer services be easily recorded. This information provides new clues for enterprises to predict the churn status of customers in advance. This work takes the online education industry as an example. Based on the customer-customer service dialogue data, we have extracted various topics and the sentiment status contained in their communications through text mining methods. Furthermore, we explore the impact of dialogue on customer churn status by using a logistic regression model. Results show that, customer-customer service communication information can significantly affect the future churn status of customers, and can help improve the prediction accuracy. Therefore, to maintain a good business relationship with customers and retain customers, enterprise managers are suggested to pay attentions to the quality of customer service communications, stand from the perspective of consumers, and better tap consumers' consumption will.

Key words: Customer Relationship Management; Customer Churn; Customer Service; Renewal Prediction; Text Mining

引言

客户流失（customer churn）是客户关系管理（customer relationship management, CRM）中的重要问题之一[1]。鉴于客户是企业的重要资产，因此研究客户流失问题对企业提升盈利能力意义重大。如果企业能够识别具有潜在流失风险的客户并进行预警，就能够采取各种挽留策略来挽回客户，从而维护公司利润。因此，对流失客户的提前预测和挽留也是企业的首要任务[2]。

传统的客户流失预测研究大多从客户的基础信息以及业务参与信息中挖掘预警因子。例如，李季等（2020）在研究电信行业的客户流失预警问题时采用了用户的通话时长、使用流量、套餐金额、服务合约等因素[3]，Jin（2022）同样聚焦于电信行业的客户流失预警问题，但是重点考虑了用户过度花费、对优惠激励的回应对用户流失的潜在影响[4]。值得注意的是，大多服务行业当前以一对一的形式为客户提供匹配客服的长期服务。基于一对一交流，客服人员对所服务客户具有深刻的了解，这些能为企业提前预判客户是否流失提供参考。与此同时，在一对一交流中，客服人员能够直接影响客户的业务体验，进而影响其继续使用服务的意愿。随着信息技术的发展，数据记录手段的提升，客户-客服通过企业 APP 交流沟通的文字对话被详细记录，这些丰富的数据为研究客户流失原因提供了重要资源。因此，从客服-客户对话信息角度挖掘预警因子可以为客户流失管理提供重要信息参考。

综上，本文将以某素质类在线教培公司的客户基本信息与客户-客服对话文本数据为基础，通过主题模型和情感分析进行文本语义信息的提炼，并结合逻辑回归模型进行客户流失预测，以期提高客户流失预警的准确性。

1 相关研究评述

研究客户流失问题的重要价值在于通过挽留潜在流失客户来延长客户的生命周期。如果能精准识别客户消费意愿变动、提前预警客户的流失可能，企业将能在潜在流失客户正式提出服务终止前，主动提供可以提升客户留存意愿的一系列策略来进行挽留[5]；或者有效规避因某些特殊原因（如客服执行能力不同）导致的与客户交流中的负面效应，并防止服务因沟通问题即刻中止[6]，从而延续客户的服务需求。

既往文献对于客户流失的研究主要借助如下四类信息进行分析：客户基础信息数据，主要由客户注册时登记的数据构成[7-9]；客户业务参与数据，主要由客户历史消费的日期、规模、消费结构、与企业的互动及他人的业务参与情况等构成[10-12]；企业营销活动数据，引入

企业的促销、个性化政策，然后观察消费者对此做出的续费决策变动[13-17]；客户业务评价数据，通过问卷、在线平台或邮件等方式收集客户对业务的意见反馈，能够捕捉其中的情感信息、流畅度、主观性等因素[18-22]。

除了上述研究提供的从客户视角出发的因素外，客服做为直接接受客户咨询、以客户为导向的企业代表群体，也是很重要的一环[23]。从客户角度看，客服的服务态度、服务模式等因素直接影响续费决策；从企业角度看，客服做为客户的直接联系人是管理体系中对客户了解最为深刻的群体，能够提供协助客户流失警示的有效管理元素。根据已有的理论研究，客户-企业关系可以分为开发期、接触期、确立期、成熟期、反复期和消退期六个阶段[24]。结合客户生命周期，客服主要在确立期、成熟期、反复期和消退期以公平、对称的信息交流和沟通心态，主动提供质量较高的服务，及时提供信息反馈与协商合作，在双方关系进入低谷时主动降低利润、让步挽留。

以往针对客服的研究问题大多围绕如何提供客服服务、提升客服质量和满意度等角度展开。例如，从开展客户服务的角度看，在服务日常维护流程中，客服需要重视沟通模式、寻求与客户建立长期良好的关系。比如 Mousavi 等（2020）对美国四大电信服务商的客户在 Twitter 上与客服的沟通进行研究，探究了各公司的客服在客户情感变动上的捕捉效率，并探讨了良好客服服务的组成因素[25]。赵卫宏等（2015）研究中国的在线零售服务业发现，在出现购物差错的情况下，客服在与客户沟通进行补救的过程中如果能反应及时、补偿到位，能够提升客户的满意度和信任度[26]。在客户流失问题的研究上，客户-客服沟通的有效性也被广泛验证。随着互联网技术的发展，企业以客户-客服沟通作为主要管理手段，同时结合各类社交媒体平台（如自有 APP 或微信等常用社交平台）和传统的联系方式（如电话、短信）面向客户。相较于传统的、企业主导的、静态的客户关系管理，这种社会化的客户关系管理赋予了客户互动、沟通的权利，强调与客户协同、联动，能够形成一种双向的良性价值共赢。

由于客户-客服沟通的内容通常以文本的形式留存，因此通常采用文本分析方法进行研究。目前已有研究利用文本信息进行客户流失预测。例如 Coussement 等（2009）和夏国恩等(2018)关注客户发出文本信息中的情绪对客户流失情况的影响[20-21]。Sahni 等（2018）采用实验的方法，通过在电子邮件中设计特定的文本信息来研究邮件内容对客户是否订阅的影响[16]。Caigny 等（2019）使用 LSTM 模型从客户与理财顾问的沟通信息中挖掘特征，并研究其对客户流失的影响[27]。Vo 等（2021）将客服与客户电话语音转成文本，然后构造词语重要性指标等特征来研究用户流失预测问题[28]此外，为了更好的理解文本内容，LDA（Latent

Dirichlet Allocation）主题模型也是一种常用方法[29]。该模型的核心思想是采用一种无监督的方式历练沟通文本中蕴含的各个主题，从而对客户-客服沟通的内容进行浓缩和概括。基于主题模型研究客户-客服沟通内容的相关工作较多。例如，Slof 等（2021）基于某电信服务提供商的客户-客服通话记录的文本数据，使用 LDA 提取文本中的主题作为自变量，发现包含主题变量的模型可以产生最佳的流失预测效果[30]。Kwon 等（2021）基于医疗公司的生活日志和用户短信数据，通过主题模型挖掘文本特征并预测数字医疗保健用户的流失情况[31]。与以往文献相比，本文将同时使用主题模型和情感分析来挖掘客户-客服对话数据，并且根据实际应用场景中的业务知识设计了自定义主题，在分析对话文本的基础上兼顾行业经验，从而获得更加全面的分析维度。

2 理论分析与研究假设

2.1 互动理论及其研究成果

在客户流失研究问题中，客户-客服沟通具体的表现为客户与业务方的管理人员通过 APP、电话等进行文字、语言交流。不难看出，客户-客服沟通可以定义为一种客户与业务方之间的互动，对该互动的体验对于客户的留存意愿有直接、深刻的影响。因此，下面将着重从互动理论的角度阐述客户-客服沟通对客户流失管理的贡献。

互动理论发源于社会心理学家舒茨（W.Schutz）在 1959 年提出的人际关系三维理论（Three Dimensional Interpersonal Relations Theory）。该理论认为：每个处于社会中的个体都具有人际交往的愿望与需要，并从包容、支配和情感需要三方面加以概括[32]。因此，人际关系的实质是双方在心理和行为两方面的双重互动，需要双方同时表现出包容、友善、信任的情感，以期维护社会关系的长期、稳定[33]。

从互动理论出发对客户-客服沟通进行辨析可知，对于客户与客服两个个体建立的互动关系，更加要求双方，尤其是客服一方，主动扩大对方对自己的认知，加深对自己的信任。这是因为这类关系的建立较为特殊：是一种无视双方人际吸引、随机匹配而开始的关系，双方以维护业务交流为核心，且客户具有单方面中止互动关系的权利，无需承担任何人际冲突乃至人际破裂的成本。

从对已有研究的梳理来看，良好、稳定的互动关系能够对客户产生积极影响。McMijlan（2005）认为感知价值对顾客的购买意愿有着积极的影响，而这种感知的形成来源于积极正向的互动体验[34]。Yen（2011）经过研究发现，互动的响应性和双向性有助于提升客户的业

务满意度和粘性[35]。此外，汪旭晖等（2015）[36]与韩雨彤等（2022）[37]对直播电商服务中的客户-主播交流进行研究发现：主播对直播间客户的提问回复能拉进与客户的心理距离，建立与客户之间的情感连接，营造更加轻松、舒适的购物氛围，由其他客户组成的活跃互动氛围也使新进客户更容易和主播乃至整个直播间群体产生心理和情感上的连接，满足了消费者自身的归属需要和社交需求。孟庆斌等（2019）对股市交流平台上目标企业与投资者沟通的内容与股价崩盘风险之间的关系进行了研究，发现沟通内容能够降低股价崩盘风险 [38]。卞世博等（2022）也证实了投资者与上市企业在平台上的高质量内容有助于投资者成为对应企业的长期注资人[39]。

通过上述论述，为了维护客户关系、防止客户流失，客服方应在互动过程中改善互动方式、使用互动技巧，以期改善客户对双方互动的感知，进而提升其对于客服的信任度和留存意愿。

2.2 互动理论在客户-客服沟通场景下的应用

为了向客服群体提出在与客户互动时的注意事项，将互动理论转化为具体的指导性操作建议，下面将在客户-客服沟通互动场景下进行讨论。

在互动过程中，沟通内容是业务导向的人际关系中的重中之重，其着眼点在于客户对服务质量的感知。Parasuraman 等（1985）提出了感知服务质量差距模型，认为顾客感知服务质量决定了顾客对服务质量的评价，而顾客感知服务质量取决于服务过程中顾客的感知与顾客对服务的期望之间的差异程度[40]。刘俊清（2018）则直接指出，感知价值与客户满意度具有密切相关、相互补充的关系，且感知价值的高低会直接影响客户满意度[41]。在客户-客服沟通场景中，双方的互动多是通过手机 APP、企业微信等方式开展，这种互动与客户单独面对机器客服不同，不是满意度调查、投诉处理等客户单方面提交业务错漏的高度机械化流程，而是一种结合客户导向的发展理念，在满足顾客针对业务提出的疑问时不仅着眼于解决问题，更要深化同客户互动的关系、尽量与顾客感知的服务质量保持一致，最终为保持双方人际关系和客户对业务的满意度而采取的一系列措施[32]。于是本文提出如下假设：

H1：和顾客使用产品有关的正向积极互动可以减少客户流失。

礼仪准则是在互动过程中，除了互动内容外最重要的一项内容。英国学者 Leech（1983）将礼貌原则具体阐述为六条准则，分别为：得体准则（Tact Maxim）、慷慨准则（Generosity Maxim）、赞誉准则（Approbation Maxim）、谦逊准则（Modesty Maxim）、一致准则（Agreement Maxim）和同情准则 （Sympathy Maxim），这六条准则体现了人际交往中避免冲突的核心

原则[42]。以这六条准则进行沟通时，客服应能够充分预设客户的需求、投诉、抱怨，让客户感受到业务方对客户利益的维护与考虑，在展示服务专业性的同时，提升客户购买服务或续费留存的意愿。同时，客服可以以较为亲密的态度维护与客户的人际关系，如近年来各电商平台常使用“亲” 等称呼或表情符号拉近与客户的距离感[43]。在互动过程中，客服友好的言语对顾客而言往往意味着善意和尊重，在一定程度上能满足顾客在双方互动时的亲密与包容需要。于是本文提出如下假设：

H2：互动过程中客服的礼貌用语可以减少客户流失。

此外，客户在互动过程中的情感也具有一定的提示作用。在一定程度内，客户的话语情感越饱满，越能体现客户对服务的满意度，因此预示着客户越有可能续费。然而，超越这个范围后，过度饱满的情感则可能体现了客户在与客服交往中的"社交奉承"或"社交称赞"的心理。具体来说，在社交关系中，人们可能出于礼貌或为了维护良好的人际关系而选择说出符合他人期望的话语，或者为了避免伤害对方的感情或引起冲突而表达一种与他们真实感受不符的态度。既往文献已经对“社交奉承”现象进行了诸多研究。例如：Chan 和 Sengupta（2013）着眼于销售人员对顾客的社交奉承心理。他们通过设计实验发现，即使是被认为是真诚的奉承，顾客也会对奉承者产生自动的负面反应。Li 等（2016）运用主成分分析和结构方程模型的方法研究了不同类型的用户奉承与产品复购意愿的关系。他们发现用户的真实赞美对再购买意愿有正向影响，而刻意表扬对再购买意愿具有负向影响。Danziger（2020）给出了社交奉承的定义并且系统的研究了如何判断一个行为是不是社交奉承。他们认为，奉承是一种明显的交际行为;它的目的是为了取悦接受者，这种效果调解了奉承者的三个互动目标之一:交易、自我推销或社交关系。他们发现，除了赞美之外，感谢、承诺、问候、道歉等都有可能是奉承行为，并且判断是否是奉承时需要对奉承者意图进行评估。

在本文的场景中，客服向客户推送新的续费促销活动时，部分已经坚定不续费的客户往往将首先肯定业务质量，再以个人安排受限为由推辞续费，并用礼貌的言语对客服进行感谢。这类对话的情感往往也呈现出高度的正面性，这体现了客户的“社交奉承”心理。但是，这些赞美和肯定并不代表客户具有续费意愿；同时，虽然不续费的理由不一定是真实的，但是客服应当注意到客户确实提出了不续费的诉求。从客户的角度，这种表达方式将不续费的理由归于自身的客观情况，对业务质量和客服的服务效果进行了免责声明，能在一定程度上缓解客服的沟通压力，有利于维护社交关系。于是本文提出如下假设：

H3：互动过程中客户的情感态度和客户流失具有非线性关系。

此外，客户-客服的互动存在一类较为重要的特殊节点：在特殊的促销活动上线时，客服会采取互动营销策略。互动营销是指在借助互联网信息技术的优势下，实现移动通信网络和互联网的有机融合，通过创建移动终端设备为载体，定 位精准的消费群体，实施营销目标和营销策略的营销方式[44]，客户对互动营销沟通的态度比单向传播更为积极 [45]，有效的营销沟通方式可以从积极方面引导客户的的心理依恋，提高其忠诚度[46]，业务方也能借此通过与消费者的互动，培养独属于自己的忠实客户群体，从而达到精确定位的目的[47]。具体到客户-客服沟通场景，业务方主要的互动营销策略为：配合促销活动，针对全体客户发送促销信息。在这种场景下客服不仅承担售后服务的角色，也承担营销人员的角色。

事实上，这类互动营销策略并不一定能够为业务提供助力，反而对业务方来说是一把“双刃剑”：如果能够结合客户的日常业务参与行为，实现精准的个性化互动营销策略，将有极大的概率吸引客户主动询问与参与这类促销活动；反之，如果采用了机械、粗糙的或者个性化程度不足的促销信息，反而可能引起客户的反感和无视，无法达到预期的沟通效果。于是本文提出如下假设：

H4： 精准营销互动可以减少客户流失，但机械化营销互动会加速客户流失。

3 研究框架

为了研究基于客服-客户文本对话信息的客户流失问题并研究上述假设，本文以某互联网素质类在线教培公司开展的付费业务为例进行分析。该付费业务主要针对学龄前儿童开展益智娱乐服务，以趣味动画和互动游戏为主要业务内容，目的在于培养儿童动手、归纳等基础思维能力。该公司为客户提供的各类服务均有一定期限，到期前需要客户续费以便继续使用该服务。因此，客户流失问题在该公司的实际业务场景中具体表现为客户续费问题。本文将重点关注客服人员与客户方家长进行的对话数据，对话主要围绕业务续费、时间调整、服务内容等一系列主题展开。同时，客服人员会基于客户的服务参与情况和服务到期时间，主动提醒客户考虑续费问题，并有探究客户续费意愿的实际需求。

在与客户沟通交流方面，客服人员主要采用公司 APP 中内置的文字交流平台进行。客户的课程时间安排、权益兑换、孩子课堂表现、续费诉求等业务需要由人工客服通过该内置交流平台与客户进行文字交流沟通。除此之外，客服人员可以通过后台查看客户填写的注册身份信息、日常 APP 使用行为、孩子课堂与课后活动参与表现等数据，以此作为与客户文字交流的补充。值得注意的是，对于所关注的在线教培行业客户个体，其续费诉求目前普遍

由人工客服通过与客户文字交流实现。因此，基于客户与客服对话的文本分析对于预测该教培付费业务的持续性有较为重要的意义，也是目前客户关系管理研究中重要的方向之一。

以上述业务背景为基础，为了探索客户-客服沟通信息对客户流失的影响，本文将重点开展两方面的研究内容。第一，采用文本主题模型，从客户-客服对话文本中提取有效信息。面对长期一对一交流的客服人员，客户在续费意愿的表达上往往更为直接、明了，因此在客户-客服对话文本中可以捕捉到客户续费意愿的变动发展。为了捕捉这些信息，本文将采用文本挖掘的经典方法——主题模型，试图从对话文本中提取出有意义的主题信息。第二，采用逻辑回归模型探索提炼出的主题信息对客户流失问题的贡献。为此，研究中将通过纳入控制变量、进行模型对比的方式，探究增加客户-客服对话文本主题信息后，流失客户的预测准确率是否会显著提升。图 1 展示了本文的具体研究框架。

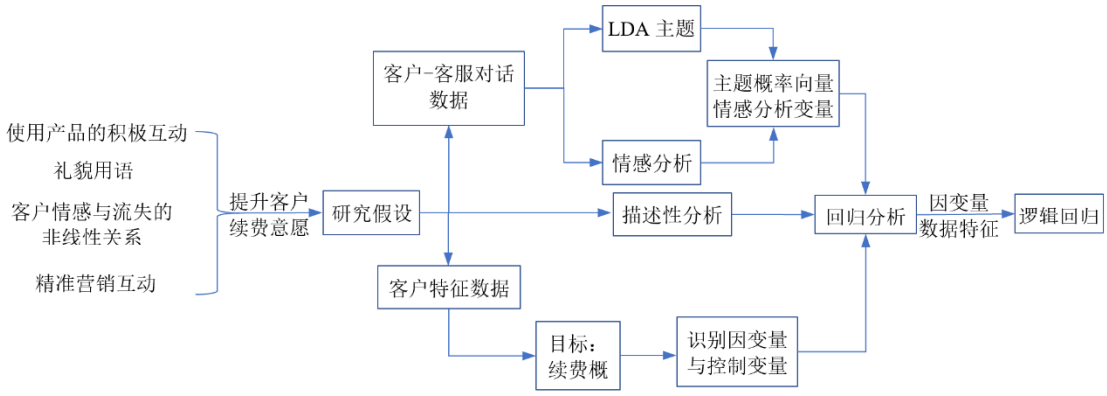

图 1　研究流程示意图

与既有文献相比，本文的主要贡献在于如下两方面。首先，在理论上，本文结合教培行业的业务特色对营销互动进行了研究。文中将营销互动分为精准营销互动与机械化营销互动两类，并分别研究它们对客户流失的影响。此外，文中也研究了基于产品展开的积极交流与客服礼貌服务用语这两个维度对客户流失产生的影响，以及客户情感与客户流失之间的非线性关系。这些研究工作丰富了当前客户-客服研究的理论框架。（2）以客户-客服对话文本数据为重点研究对象进行流失因子挖掘，在充分了解业务流程和服务内容的前提下，采用数据驱动的主题模型、情感分析等文本挖掘方法，提炼归纳出客户群体的评价主题和沟通交流中表现出的情感态度，然后进一步研究主题和情感对客户流失的影响，从而验证理论假设的正确性。通过这些实证研究内容可以加深对客户业务参与行为的深刻理解，从而帮助业务方改善业务流程和客服管理，降低客户流失率。

4 数据说明

本文将采用两部分数据：（1）客户特征数据，包括客户是否续费以及其他行为特征，（2）客户-客服对话数据，以文本形式存在。下面将对这两部分数据进行介绍。

（1）客户特征数据

客户特征数据是根据该公司 APP 所记录的客户行为而定义的一些客户基本特征，包括 43,969 个客户在 2021 年 11 月的特征信息，包括客户历史的续费情况、孩子的基础情况、课堂参与表现等共 64 个特征；以及这些客户在 2021 年 12 月份的最终续费状态。表 1 列出了部分客户特征字段的具体情况。本文的因变量为是否续费（is_renew）。在 12 月涉及的 43,969 个客户中共有 8,029 个客户于本月最终续费，整体续费率为18.26%。

表 1　客户特征数据内容举例说明

类别		字段
因变量		是否续费（is_renew），为 0-1 变量
自变量	客户 消费记录	客户历史购买单量（user_order_num） 客户首次付费距该月月份数（first_order_month_interval） 客户月初剩余课时数（class_hour_left_total） 客户月初剩余代币可兑换课时数（student_left_valid _coin） 续费券是否本月过期（coupon_end_current_month） 客户在 12 月的续费日期（pay_date），如未续费则为空
	客户业务 参与表现	该月月初过去 45 天作业完成率（homework_finish_rate） 该月月初过去 45 天测评参与率（phase_join_rate） 该月月初过去 45 天课时消耗数（class_consumption）
	客户 情况数据	城市等级（city_level_dummy_新一线城市） 分班等级（class_level_dummy_A+） 客户质量等级（allocation_level_dummy_B2）

（2）客户-客服对话文本数据

客户-客服对话文本数据来自于公司开发的 APP 软件以及一些常用的社交平台，每个客户可以与负责自己孩子培训的客服一对一联系。所有对话文本数据共有 1,258,859 条，涵盖了 62,704 个客户、1,064 个客服在 12 月产生的聊天记录。表 2 展示了 3 条对话文本数据（为保护商家数据隐私，示例中信息均为脱敏处理）。一条数据都附有沟通双方 ID，对话发送者和发送时间、聊天的具体内容和沟通所使用的方式。

表 2　对话文本数据示例

客户 ID	客服 ID	消息类型	消息产生时间	内容	沟通方式
123	456	客服发送	2021-12-11 12:34:56	鹏鹏妈妈您好，这边技术反映孩子说课件显示有问题，请问需要帮助吗	自有 APP
123	456	客户发送	2021-12-11 12:34:57	需要的	自有 APP

123	456	客服发送	2021-12-11 12:34:58	好的，请妈妈稍等哈！这边马上联系您	自有 APP

5 客户-客服对话文本分析

5.1 主题模型方法介绍

以 LDA (Latent Dirichlet Allocation) 模型为基础的主题模型是文本分析中一类非常重要的模型，在近年来引起了学者的广泛关注。LDA 模型假设文本中隐含着丰富的话题。在这种假设之下，模型可以计算每篇文档的主题概率分布，同时得到每个主题的相关信息，自动实现对文本内容的总结概括。为此，本文将使用 LDA 主题模型来提炼挖掘客户-客服对话数据中蕴含的丰富信息。

具体来说，LDA 模型假设整个客户-客服对话数据集中存在 K 个潜在主题，每个对话在既定的 K 个主题上存在潜在概率分布，而每个主题也存在一个概率分布，这样就构成了“对话-主题-词”的层级结构概率模型。在 LDA 模型中，定义$\boldsymbol{\theta}_i = (\theta_{1i}, \dots, \theta_{Ki})^T \in \mathbb{R}^K$为第$i$句对话在所有 K 个主题上的概率分布参数，$\boldsymbol{\phi}_k = (\phi_{1k}, \dots, \phi_{Vk})^T \in \mathbb{R}^V$为第 k 个主题在 V 个词所构成的词典空间上的概率分布参数。进一步，定义α为$\boldsymbol{\theta}_i$分布的超参数，定义β为对应$\boldsymbol{\phi}_k$的超参数。为此，第i句对话的具体生成过成如下：

（1）根据 Dirichlet 分布以及超参数α产生第i句对话在所有 K 个主题上的概率分布，即$\boldsymbol{\theta}_i \sim Dir(\alpha)$。

（2） 第i句对话中的第 n 个词按照如下步骤产生：

（2.1） 根据多项分布，首先产生该词所表达的具体主题：$z_{in} \sim Multi(\boldsymbol{\theta}_i)$；

（2.2）根据多项分布，产生表达该主题的具体的词：$w_{in} \sim Multi(\phi_{z_{in}})$。

（3） 根据 Dirichlet 分布和超参数β，产生主题 k 的概率分布$\boldsymbol{\phi}_k$~Dir(β)

按照上述生成过程，可计算所有对话的联合似然函数并进一步求解得到参数$\boldsymbol{\theta}_i$和$\boldsymbol{\phi}_k$的估计值。通过$\boldsymbol{\phi}_k$，可以知道每个主题下出现概率最高的词，从而概括这一主题的含义；通过$\boldsymbol{\theta}_i$，可以知道第i句对话在不同主题上的讨论权重，这一权重将作为特征字段用于预测客户的流失状态。

在使用 LDA 模型时需要事先确定主题数 K。通常可以采用一些评价指标，并结合主题模型的效果来选择 K。本文中采用两个常用的指标来选择 K。第一个指标为困惑度（Perplexity），它经常用于衡量语言模型的好坏，其原理是根据每个词估计一个完整句子在文本中出现的概率。困惑度越小，语句在文本中出现的概率越高，说明模型的效果越好。第

二个指标是一致性得分（Coherence Score）。该指标衡量了主题中概率较高的词在实际文档中同时出现的概率大小，更高的一致性分数表示主题具有更好的解释性和语义上的连贯性。

5.2 主题模型结果

首先采用 Jieba 对所有的对话文本数据进行分词。为了提升分词效果，引入停用词典和保留词典进行文本分词的优化，这也是中文文本处理的常见操作[48,49]。在构建这两个词典时，我们首先人工阅读部分客户和客服的聊天记录，然后专门针对语料的口语化问题和专有名词遗漏问题设计而成。这两个词典的使用，可以帮助 Jieba 捕捉到许多与续费问题高度相关的词语（如专有代币名称、优惠券名称）。同时，对于已经在当月续费的家长，将根据续费日期将续费后的语料切除。

对于分词后的文本数据，采用 LDA 主题模型进行建模。在建模时，采用困惑度和一致性得分来确定最优主题数 K。主题数 K 的筛选范围从 1 到 15。建模结果显示，困惑度会随着主题数的增大而不断降低，说明困惑度更倾向于选择更大的主题数；一致性得分在主题数较小时取值较高，随着主题数的增大略有下降，说明一致性得分更倾向于选择更小的主题数。两者结合，并通过人工判断各个主题的含义，最终确定主题数 K=11，并以此进行后续的文本主题分析。表 3 展示了每个主题中概率最高的前 10 个词，据此可以对主题进行命名。可以看到 11 个主题均呈现出清晰的语义，说明主题提取的效果较为良好。下面将结合各个主题的前 10 个高概率词具体分析。

表 3 主题模型结果

编号	主题命名	高概率词
1	客服礼貌用语	爱心、打扰、回复、忽略、测评、真的、错过、拥抱、我哈、学员
2	促销期：预设消息	愉快、机智、元旦、宝爸、级别、庆祝、礼物、年终、保障、快乐
3	日常：预设消息	编程、会员、羊毛、学具、邮寄、基础、流泪、历史、班主任、星球
4	促销活动	海报、直播间、抽奖、即可、直播、二维码、参与、续课、预约、儿童
5	课程协调	请假、补课、生成、一年、沟通、加油站、好礼、手机号、回放、题目
6	促销期：续费	京东、优惠、十二、领取、下单、续费、领券、满减、优惠券、专属
7	日常：续费	代币、课时、福利、购课、兑换、赠送、续费、十二、课包、年月日
8	学生测评	参加、运动会、自评、报告、主会场、规划、第八届、本次、未来、收获
9	课后巩固提醒	截图、秘籍、培养、能力、跳级、复习、小游戏、阶段、特别、课后
10	朋友圈打卡	打卡、点击、朋友圈、自评、参与、方式、加油、图片、步骤、右上角
11	业务提醒	发放、家人、抱拳、希望、录播、提交、链接、一栏、益智、截止

主题 1 是以家长和客服的日常沟通为主的主题。这个主题以客服话语为主，涵盖了很多礼貌用语如“爱心”、“打扰”、请“忽略”等，主要的目的是提醒家长不要错过孩子的线上测评

机会，以考察孩子最近的业务参与状况。基于主题高概率词的含义，将主题 1 命名为“客服礼貌用语”。

主题 2 主要呈现客服向家长反映孩子年底活动参与表现的对话内容，是促销期间系统预设的群发消息。客服会在年终时刻祝客户元旦愉快，并庆祝孩子完成了本年度的业务参与、适时地为孩子邮寄小礼物，以促进双方的沟通磨合情况。基于主题高概率词的含义，将主题 2 命名为“促销期：预设消息”。

主题 3 以家长和客服的日常沟通为主，展示了客服在各类日常活动中发送的预设消息内容。具体而言，一方面是新业务的尝试触达，如编程、历史课程；另一方面是日常学务活动的沟通，如学具的邮寄、与班主任老师的沟通交接等。基于主题高概率词的含义，将主题 3 命名为“日常：预设消息”。

主题 4 反映的主要是该月公司举办的一次抽奖活动。家长可以以宣传海报上注明的报名方式预约参与，到时间通过公司发送的直播二维码进入直播间，即可参与公司的年底感恩回馈抽奖活动；在直播间下单续课还会有额外的优惠赠送。基于主题高概率词的含义，将主题 4 命名为“促销活动”。

主题 5 则反映了客服与家长在课程协调上的交流和反馈。其中，“请假”、“补课”一般是家长方提出的诉求；“生成”“一年”的学情报告、参与“加油站”好礼兑换、查收课程“回放”和课后“题目”等一般是客服人员对客户业务内容的督促。基于主题高概率词的含义，将主题 5 命名为“课程协调”。

主题 6 是研究中最关注的主题：续费情况询问。客服一般以账户内剩余课时数量较少，或账户内优惠券即将过期的客户为目标群体，向家长反映系列问题，用新到账优惠券或课包购买优惠为理由，询问家长是否领券并下单、是否有足够的续费意愿。在 12 月这个特殊时刻，公司以京东作为合作活动平台，建议家长在有满减折扣的日期参加续费活动，并适时地提醒家长在规定时间内创建订单、完成支付。基于主题高概率词的含义，将主题 6 命名为“促销期：续费”。

主题 7 则反映了日常的续费情况询问问题。此时的续费意愿涉及的双十二相关内容较少，主要是以家长账内特殊代币数量较少、课时余额不足，同时此刻购课存在兑换/赠送某些实体礼品或账户课时的福利时，作为询问是否续费的好时机。基于主题高概率词的含义，将主题 7 命名为“日常：续费”。

主题 8 涉及公司举办的一次规模较大的测评活动“运动会”，公司会针对参与的孩子水平进行会场划分，参加的孩子会以自评测验的形式进行竞技，再在结束后生成报告收获本次测评的结果。基于主题高概率词的含义，主题 8 命名为“学生测评”。

主题 9 主要着眼于公司对于客户在业务参与结束后的巩固活动提醒。其中，“秘籍”是公司编写并会发送给客户的一些益智习题，并鼓励家长带着孩子练习、有疑问可以发送截图询问客服；除了“秘籍”以外，为了督促孩子的能力培养，客服会经常提醒家长以包含做游戏在内的方式督促孩子进行复习活动。基于主题高概率词的含义将主题 9 命名为“课后巩固提醒”。

主题 10 以家长分享自己的参与情况以吸引更多潜在客户与公司联系、缴费为主。家长在督促孩子结束学习后，可以参与自评；在自评结束后，通过点击右上角的分享按钮可以生成打卡图片，在朋友圈分享后截图发给客服，以获得一些额外的分享福利。基于主题高概率词的含义，将主题 10 命名为“朋友圈打卡”。

主题 11 主要着眼于客服对客户业务参与情况的提醒。其中，“发放”是客服给客户优惠券时的常用话语；除了“发放”以外，为了督促孩子的能力培养，客服会经常提醒客户督促孩子提交课后练习、及时完成益智游戏等。基于主题 11 高概率词的含义，将该主题命名为“业务提醒”。值得注意的是，在后续模型建立的过程中，将略去主题 11 的概率主题向量以回避多重共线性。

可以发现，上述 11 个主题刻画了客户和客服之间沟通的主要内容，包括家长需要进行课程协调，客服发送的课后巩固提醒，促销信息等。同时注意到，这 11 个主题并没有反映客户对教学使用的满意度或者细节反馈等内容。这可能是由于该公司提供的课程是一种标准化的产品，而非用户定制产品。大多数客户会通过试听的方式了解该产品的情况，如果合适会直接购买，不合适就直接拒绝，因此在与客服的沟通过程中对产品或者服务进行评价的内容较少，以至于主题模型并没有提炼出相应的主题。为了弥补这一缺憾，本文根据客户-客服对话中各个词语出现的词频，并结合行业经验和公司业务知识，设计了一个新的主题“教学内容反馈”，并编号为主题 12。支撑该主题的词语为“答案, 表达能力, 难度, 提问, 回答, 耐心, 难, 细心, 重复, 贴心, 练习, 简单, 效率, 思考”。可以看到，这些词语反映了用户（即学生）上课时的表现、对课程的感受，以及授课教师的表现。每个用户在该主题上的表达概率即为用户发出的所有对话内容中这些词语出现的总频率。

5.3 情感分析

客户-客服对话文本中往往蕴含着一定的情感信息。如果在沟通的过程中客户表现出较为强烈的正向情感倾向，可能表达了客户对课程服务较为满意，因此客户可能会有较高的

续费意愿。但有时，客户也会在决定不续费之时表达“善意的赞美”。所以客户情感与客户的续费意愿可能具有较为复杂的非线性关系。为了探究这一假设，本文使用 Python 中的 Snownlp 工具对客户发出的对话内容计算其情感得分，该得分在 0 到 1 之间，得分越高表示情感越偏正向和积极，反之得分越低表示情感越偏负向和消极。进一步，对每个客户发出的所有对话文本的情感得分取平均值，作为该客户的最终情感得分，没有说话的用户情感得分取为 0。图 2 展示了所有客户的情感得分分布直方图。从中可以看到，绝大多数用户的情感得分在 0.6 分以上，整体较为积极。

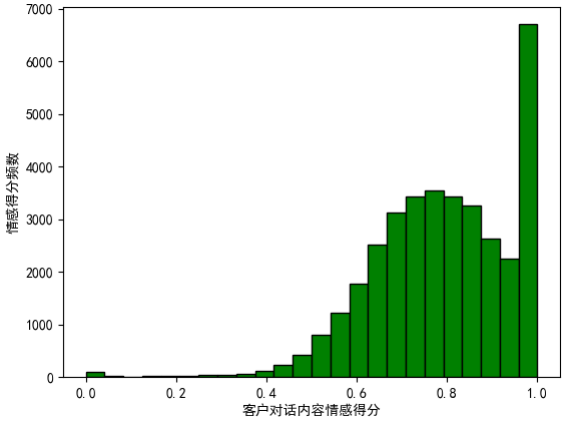

图 2 客户对话内容情感得分分布直方图

为了进一步探索客户情感态度与续费率之间的关系，我们将用户情感得分按照三分之一和三分之二分位数划分为三组，并分别计算每组人群的续费率，发现按照情感得分由低到高的三组续费率分别为：15.8%、19.2%、16.2%，这可能暗示着续费率与情感得分呈现非线性关系。这种非线性关系来源于两方面：首先，在一定程度内，客户的话语情感越饱满，越能体现客户对服务的满意度，因此客户越有可能续费。然而，超越这个范围后，客户可能出于“社交奉承”的心理[50-52]，仍然对客服和产品服务表达非常积极的情感，但是确没有续费意愿。图 3 展示了部分典型的客户对话，由此可见，用户在表达赞美、感谢的同时，其实并没有很强的续费意愿。在后续研究中，我们会将客户情感得分的一次项和二次项同时作为解释变量放入续费模型中，以探究客户情感态度和客户流失之间可能存在的非线性关系。

客户ID	客服ID	消息类型	消息产生时间	内容	情感得分
757235	418	客户发送	2021/12/17 20:03	茶茶老师谢谢您我们没有课时了暂时也不准备再续课非常感谢您这几年的教导	0.983
737499	1097	客户发送	2021/12/22 20:39	谢谢罗老师一直以来尽心尽责的陪伴你真的是一位很负责任很有爱心脾气很好的老师祝你和宝宝健健康康平安喜乐	1.000
11004	2471	客户发送	2021/12/3 15:47	太感动了谢谢三只鱼老师玫瑰玫瑰玫瑰我跟严溯也会很想念温柔美丽耐心有爱的三只鱼老师的也会跟严溯一起努力成长希望以后有机会续课能再跟老师见面爱心爱心爱心	1.000
446108	2524	客户发送	2021/12/23 21:00	谢谢潘老师这么长时间以来劳烦你费心了我们以后会继续加油的感恩爱心后续如有有需要再来麻烦你握手	0.998
645910	4564	客户发送	2021/12/28 7:24	谢谢张老师的耐心点评通过你对糖果这次测试的点评我知道哦糖果哪些地方的不足今后我会带她多练习谢谢	0.999

图 3：部分高情感得分的客户对话内容

6 建立客户流失预警模型

6.1 模型设计

下面对所采用的数学符号进行陈述并给出模型设定。用Y表示因变量：$Y=1$表示客户续费，$Y=0$表示客户不续费。用$X=(X_1,X_2,\dots X_{66})'$定义用户特征信息表中除用户 id 与“是否续费”以外的自变量和衍生开发的情感变量，共 66 个变量。对于主题相关的自变量，通过 LDA 模型可以得到每个对话文本在 K 个主题上的概率值。对于每个用户而言，需要将该用户全部的“客户-客服”对话文本在某个主题上的概率值取平均，从而得到该用户在某个主题上的平均概率值。考虑到某个用户在所有 11 个主题上的平均概率值求和等于 1，因此为了避免多重共线性，只放入前 10 个主题的概率值。除此之外，对第 12 个人为设定的主题，也依据类似的方式，计算每个用户发出的所有对话内容中特定词语出现的总频率，以此作为该用户在这个主题上的表达概率值。通过上述方式可以得到 11 个主题概率向量，定义为$X_{\text{topic}}=(X_{t1},X_{t2},\dots X_{t10},X_{t12})'$。与此同时，定义$X_{\text{senti}}$为客户情感得分的一次项和二次项。除了上述变量的原始取值之外，本文还额外考虑了变量的幂次效应和交互作用，下面将详细介绍。

公司的日常业务实践证实了客户账户剩余课时数（class_hour_left_total）可能对用户续费产生较为复杂的非线性影响。为了刻画这一点，本文构建了账户剩余课时数（class_hour_left_total）的二次、三次、四次项作为新的解释性变量。同时，公司的日常业务运行还发现账户剩余课时数（class_hour_left_total）和有无剩余代币（has_spark_coin）之间可能存在协同作用，通过客服人员的反馈得知：在账户的各类权益中，剩余课时数是业务核心内容的量化体现，是客户最为关心的权益指标，其数量对客户的续费意愿有较强的影响力。同时，代币作为一种可以兑换课时的特殊权益，可能与剩余课时数共同发挥作用，对客

户的续费意愿进行影响。因此，本文进一步构建账户剩余课时数的幂次项与有无剩余代币的交互项。上述所有变量定义为自变量X_{inter1}。

其次，通过客户-客服文本提炼生成的内容主题，也有可能和其他变量存在协同作用。本文主要考虑两种。其一，考虑到不同地区的客户在业务竞争意识上具有异质性，从而可能影响续费意愿，且该公司的业务约四分之一的客户来自北京市，因此北京市的客户可能具有较强的异质性。因此考虑设置主题 8“学生测评”和“客户所在地是否在北京”之间的交叉项。其二，考虑到剩余课时数对续费意愿的影响可能较大，因此尝试设置主题 6“促销期：续费”和“剩余课时”的交叉项，以及主题 7“日常：续费”和“剩余课时数”的交叉项，以此探究“剩余课时数”与其他两个主题之间是否具有协同作用。除此之外，本文还额外增加了变量“聊天记录条数”，用于衡量客户和客服之间的沟通强度，并进一步考虑了该变量和主题 1“客服礼貌用语”的交互作用。将上述四个交叉项定义为自变量X_{inter2}。

设$X_{\text{all}} = (X, X_{\text{topic}}, X_{\text{sneti}}, X_{\text{inter1}}, X_{\text{inter2}})'$为全部自变量的集合。为了研究自变量与客户续费之间的关系，建立如下三个逻辑回归模型：

$$\text{模型 1:}\ \text{logit}\{p(Y = 1|X; \beta_1)\} = \text{X}\beta_1$$

$$\text{模型 2:}\ \text{logit}\{p(Y = 1|X_{\text{topic}}; \beta_2)\} = X_{\text{topic}}\beta_2$$

$$\text{模型 3:}\ \text{logit}\{p(Y = 1|X_{\text{all}}; \beta_3)\} = X_{\text{all}}\beta_3$$

其中，$\text{logit}(x) = \log\{x/(1-x)\}$为逻辑变化，$\beta_1$、$\beta_2$、$\beta_3$表示三个模型的回归系数。同时研究中将采用 BIC 准则对所建立的逻辑回归模型进行逐步回归，以完成变量筛选，从而提高模型的整体效果。为了评估逻辑回归模型的预测效果，本文采用一系列指标来评估三个模型在测试集上的预测结果，包括：准确率、精确率、灵敏度、特异度、F1 值和 AUC 值。

6.2 模型估计结果

由于对话文本数据与客户特征数据抓取标准不同，两个数据集之间的客服、客户信息不能完全匹配。为此，在建立逻辑回归模型之前，首先匹配两部分数据并删掉字段取值为空的数据，最终共有 35,779 位客户进入模型分析。为了衡量模型的的预测效果，将匹配后的数据集随机拆分成训练集（28,624 个家长）和测试集（7,155 个家长的信息）。在训练集上建立两个逻辑回归模型，并使用 BIC 准则进行逐步回归变量筛选，三个模型的回归系数如表 4 所示。

表 4 展示了模型 3 的估计结果；模型 1 和模型 2 的估计结果见附录。从表 4 可以发现，采用了用户特征的模型 1、3 保留的用户特征信息基本一致，采用了主题因子的模型 2、3 也保留了基本一致的主题信息。总体来说，“客户历史续费次数”正显著，说明客户在历史续费次数较多后呈现出较高的忠诚度，然而“第一次付费距今月数”负显著，表面客户生命周期越

长越有可能流失。“该月月初过去 45 天完课率”、“课后练习完成率”、“测验参与率”、“测验题目正确率”、“平均课堂答题次数”、“平均课堂答题正确数”等孩子课堂表现、课后联系完成效果变量均正显著，说明孩子的良好课堂表现、学习态度的积极性能够提升家长的续费意愿；“所在地”变量的系数均显著，说明北京、深圳两地的客户相比国内其他城市更愿意续费；剩余课时数的幂次项及其与是否有代币的交叉项均入选模型，说明这类变量在探究用户的续费意愿时的确具有重要地位。

关于客户-客服对话文本信息，可以发现共保留 9 个显著的主题变量。下面将具体分析各个主题的情况并以此探究本文所设立的研究假设的正确性。

主题 1“客服礼貌用语” 体现了客服在日常业务活动中语言的亲和力，能够以一种亲近而不冒犯的风格与客户进行交流，说明了互动过程中客服的礼貌用语可以减少客户流失，与假设 H2 相吻合。

主题 2“促销期：预设消息”、 主题 3“日常：预设消息”负显著，说明它们的提升反而导致客户续费意愿的下降，其原因是这三个主题对应的客服消息为系统预设生成、群发的话语，其用语较不具有沟通感，也无法引起客户的回应，这一结果与理论假设 H4 相吻合。

主题 6“促销期：续费”和主题 7“日常：续费”正显著的，这两个主题表达了客服所发送的两种续费活动的信息，对它们的提及也能够提醒客户优惠活动的存在，从而提升客户的续费意愿。主题 5 “课程协调”、 主题 8 “学生测评”、主题 9 “课后巩固提醒”正显著，分别体现了客服在日常业务活动中对客户业务参与的关注，它们能够体现出客服对客户的督促与关怀，因而其提升能够帮助提升客户的续费概率。上述五个主题变量均体现了客服与客户紧扣业务的积极互动会提升客户的续费率，与假设 H1 相吻合。

“聊天记录平均情感得分”变量的一次项正显著、二次项负显著，这说明情感得分的影响是二次的、非线性的。具体来说，在情感得分较低时对续费是正向影响，而在情感得分较高时对续费变成负向影响。这种非线性影响与理解假设 H3 相吻合。

除此之外，主题 12“教学内容反馈”正显著，说明当用户反馈频率越高、用户对教学情况的关注程度越高时，用户的续费意愿越强烈。与此同时，注意到所构造的主题与客户特征的交叉变量并未留在模型中，说明这类交叉变量对模型的预测帮助不大。与客服的聊天记录越多，则客户的续费意愿越高；并且客服的聊天记录数与礼貌用语构成了一个显著的交叉项。对于聊天记录的探究分析发现，客户发言的情绪高涨往往预示着业务的流失，其内容以感谢客服与讨论业务为主，然而却并不代表其具有续费意愿。

表 4 模型 3 的估计结果

变量	估计值	变量	估计值
常数项	-7.751***	孩子在业务内的年级_3.0	-0.221***
客户历史续费次数	0.221***	孩子在业务内的年级_4.0	-0.241***
第一次付费距今月数	-0.021***	是否有代币	-0.261***
剩余课时数	0.211***	剩余课时数_平方	-0.021***
该月月初过去 45 天完课率	1.081***	剩余课时数_立方	0.001***
该月月初过去 45 天缺课率	-0.931***	有无代币与课时数交叉项_平方	0.011**
该月月初过去 45 天课时消耗数	0.021*	有无代币与课时数交叉项_立方	0.001*
是否有优惠券	0.911***	家长使用 APP 查看学情报告次数	0.251*
是否在班	1.111***	聊天记录平均情感得分	2.001*
课后练习完成率	0.151***	聊天记录平均情感得分_平方	-2.651***
测验参与率	0.111*	交叉项_聊天数_礼貌用语主题	0.021***
孩子综合表现水平值	0.001**	主题 1：客服礼貌用语	6.921***
测验题目正确率	0.531**	主题 2：促销期：预设消息	-2.931***
平均课堂答题次数	0.031***	主题 3：日常：预设消息	-5.801***
平均课堂答题正确数	1.191***	主题 5：课程协调	6.371***
朋友圈海报分享完成率	-0.361***	主题 6：促销期：续费	1.431***
距上次家长发消息天数	0.001*	主题 7：日常：续费	3.101***
客服消息占总消息比例	-0.301***	主题 8：学生测评	5.781***
所在地_北京市	0.391***	主题 9：课后巩固提醒	1.131*
上次续费所购课时_50	-0.291***	主题 12：教学内容反馈	0.111**

注: ***为 0.001 显著性水平，**为 0.010 显著性水平，*为 0.050 显著性水平

6.3 模型预测结果

最后观察三个逻辑回归模型在客户流失问题上的预测效果。通过三个模型可以得到测试集中每个客户的预测标签$\hat{Y}$。根据客户的真实续费情况和预测续费情况，可以评估模型效果，具体的指标情况见表 5。首先观察三个模型的 AUC 情况。可以得出，模型 3 的 AUC 值为 0.755，相较于模型 1、模型 2 在 AUC 值上有 3.7%和 6.5%的提升。这说明增加客户-客服的对话信息后，模型的预测效果提升了；同时，单独使用用户特征和客户-客服对话信息对用户续费情况进行研究，也能获取一定的效果，且使用用户特征的效果要优于客户-客服对话信息。接下来，根据 ROC 曲线所给阈值，分别对两个模型进行因变量（是否续费）的预测，并给出其他评估指标在三个模型上的取值。可以看到，模型 3 的整体效果在各个指标上均得到提升；灵敏度和特异度则分别体现了模型 3 对于续费家长群体的明确捕捉识别。

表 5 三个模型的系列检验指标

	AUC	准确率	精确率	灵敏度	特异度	F1 值
模型 1	0.718	0.675	0.327	0.653	0.680	0.436
模型 2	0.690	0.662	0.316	0.650	0.666	0.425
模型 3	0.755	0.734	0.388	0.662	0.751	0.489

7 管理实践与总结讨论

客户流失是客户关系管理中的重要问题之一，研究客户流失问题对企业提升盈利能力意义重大。如果能精准识别客户消费意愿变动、提前预警客户的流失行为，企业将能更好地服务客户、留存客户。现有的客户流失预警研究较少考虑客户-客服的对话内容，但这一信息恰恰是客户对业务诉求的直接表达，也给企业监测客户续费意愿变动提供了良好的数据基础。

研究以某互联网教培公司的客户-客服对话数据为基础，采用 LDA 主题模型有效提取聊天内容的语义信息，然后将主题概率向量作为自变量，用于提升客户续费模型的预测效果。结果显示，体现客服对客户业务参与关注度、语言亲和力、发送续费活动相关信息的主题能够提升客户的续费概率，而使用群发消息督促客户进行续费的主题则降低了客户的续费概率。后续研究可以考虑文本主题在时间上的发展联动性，引入生存分析相关概念、考察客户续费概率的动态变化，在时间维度上建立对客户的长期观察模型等。

本文采用了数据驱动的方式，通过建立 LDA 主题模型挖掘在线教培行业客户-客服对话文本中的信息，以此来探究客户与客服互动过程中的互动内容和态度是否会对客户续费产生影响。该研究框架同样适用于其他客户-客服沟通的场景，比如电商平台的客户商品评论回复场景、客户-客服售前售后咨询场景等。除了适用于不同的业务背景外，该研究框架所分析的数据也不仅局限于对话文本，还可以包括电话沟通等其他常用业务服务渠道，通过使用语音转文字的技术对电话沟通录音进行探究。与此同时，通过对客户-客服对话内容的研究，本文发现客服与客户关于产品使用展开的正向积极互动、客服礼貌的沟通用语与精准、灵活的营销互动能够帮助提升客户的续费意愿。这提示业务方可以构建一套更为规范化、高标准的客服培训机制，及时帮助客服熟悉服务内容与客户动向，学习得体礼貌的沟通方式，并用灵活的个人表达配合整体的营销策略，降低客户对促销活动的反感，与客户建立长期良好的互动关系。

本文仍然存在着一些不足之处。首先，本文使用 2021 年 11 月和 12 月的数据开展实证研究，说明客户-客服交流信息中提取的主题信息对客户续费预测的提升效果。考虑到 11 月份有双十一，12 月份有双十二，两个月份的数据各包含一次大型促销，所以在用 11 月份训练模型然后对 12 月份进行预测时，可以在一定程度上控制促销带来的影响。对于促销因素更为严格的控制需要额外补充其他时期的数据。然而受制于数据的可获得性，本文暂时无法使用其他时段的数据对结论的稳健性进行验证。第二，本文在控制变量中使用了与孩子当月

学习情况相关的静态表现变量，比如孩子综合表现水平值、作业完成率等。相比于静态变量，这些维度的动态变化也许更能反映课程是否提升了孩子的学习情况，但同样受制于数据，本文目前无法进行动态测量，这也是本文未来重要的改进方向之一。第三，同样受制于数据的限制，我们目前无法考虑客服自身对于用户续费的影响，我们将在后续研究中进一步探索。其次，本文采用数据驱动的方式测量理论假设中提出的各个维度，即通过主题模型挖掘文本对话内容，然后根据各个主题下高概率词的共同语义进行命名，将得到的主题用来测量理论上客户-客服互动中的各个维度，以此为分析起点。然而，这种命名方式是否严谨、各个主题变量是否能完美衡量理论上的各个维度，还有待后续研究进一步证实。最后，本文目前将主题模型生成的各个主题作为独立的变量放入模型分析中，并没有考虑主题之间信息的重合。因此后续研究中也将进一步考虑对各个主题进行汇总和归纳，然后探究主题类别对用户续费的影响。

参考文献

[1] TAMADDONI A, STAKHOVYCH S, EWING M. Comparing Churn Prediction Techniques and Assessing Their Performance: A Contingent Perspective. ***Journal of Service Research***, 2015, 19 (2): 123-141.

[2] ASCARZA, E. Retention Futility: Targeting High Risk Customers Might Be Ineffective. ***Journal of Marketing Research***, 2016, 55 (1): 80-98.

[3] 李季, 张帅, 周静.服务与需求的匹配度对客户流失的影响研究:基于电信行业的客户数据实验.管理评论, 2020, 32(05):192-204.

[4] JIN H F. The effect of overspending on tariff choices and customer churn: Evidence from mobile plan choices. ***Journal of Retailing and Consumer Services***, 2022, 66(2):102914.

[5] DE MATOS M G, FERREIRA P, BELO R. Target the Ego or Target the Group: Evidence from a Randomized Experiment in Proactive Churn Management. ***Marketing Science***, 2018, 37(5):685-853.

[6] BECKER J U, SPANN M, BARROT C. Impact of Proactive Postsales Service and Cross-Selling Activities on Customer Churn and Service Calls. ***Journal of Service Research***, 2019, 23(1):53-69.

[7] 罗彬, 邵培基, 罗尽尧, 等.基于预算限制和客户挽留价值最大化的电信客户流失挽留研究.管理学报, 2012, 9(02):280-288.

[8] 夏维力, 王青松. 基于客户价值的客户细分及保持策略研究. 管理科学, 2006, 19(4) : 35-38.

[9] 罗彬, 邵培基, 罗尽尧, 等. 基于竞争对手反击的电信客户流失挽留研究.管理科学学报, 2011, 14(8) : 17-33.

[10] 周静,周小宇,王汉生.自我网络特征对电信客户流失的影响.管理科学, 2017, 30(05):28-37.

[11] 李季,张帅,周静.服务与需求的匹配度对客户流失的影响研究:基于电信行业的客户数据实验.管理评论, 2020, 32(05):192-204.

[12] FERREIRA P, TELANG R, DE MATOS M G. Effect of Friends' Churn on Consumer Behavior in Mobile Networks. ***Journal of Management Information Systems***, 2019, 36(2):355-390.

[13] ASCARZA E, NETZER O, HARDIE B G S. Some Customers Would Rather Leave Without Saying Goodbye. ***Marketing Science***, 2018, 37(1):54-77.

[14] PARK C H, PARK, Y H, SCHWEIDEL D. The Effects of Mobile Promotions on Customer Purchase Dynamics. ***International Journal of Research in Marketing***, 2018, 35(3):453-470.

[15] VENKATESAN R, FARRIS P W. Measuring and Managing Returns from Retailer-Customized Coupon Campaigns. ***Journal of Marketing***, 2012, 76(1):76–94.

[16] SAHNI N S, WHEELER S C, CHINTAGUNTA P. Personalization in Email Marketing: The Role of Noninformative Advertising Content. ***Marketing Science***, 2018, 37(2):236-258.

[17] MOSER S, SCHUMANN J H, VON WANGENHEIM F. The Effect of a Service Provider's Competitive Market Position on Churn Among Flat-Rate Customers. ***Journal of Service Research***, 2018, 21(3):319–335.

[18] RACITI M M, DANAHER T S. Embedding relationship cues in written communication. ***Journal of Services Marketing***, 2010, 24(2-3):103-111.

[19] 陈明亮.生命周期不同阶段客户重复购买意向决定因素的实证研究.管理世界, 2002, (11):93-99.

[20] 夏国恩, 马文斌, 唐婵娟, 等. 融入客户价值特征和情感特征的网络客户流失预测研究.管理学报, 2018, 15(03):442-449.

[21] COUSSEMENT K, Poel D V. Improving customer attrition prediction by integrating emotions from client/company interaction emails and evaluating multiple classifiers. ***Expert Systems With Applications***, 2009, 36(3):6127-6134.

[22] 齐托托,周洵,王天梅.在线评论特征对知识付费产品销量的影响研究——基于产品类型的调节作用.管理评论, 2021, 33(11):209-222.

[23] Kassemeier R, Alavi S, Habel J, et al. Customer-oriented salespeople's value creation and claiming in price negotiations. ***Journal of the Academy of Marketing Science***, 2022, DOI:10.1007/s11747-022-00846-x.

[24] 王庆国, 蔡淑琴. 客户服务分类研究.管理评论, 2004, (03):19-24+63.

[25] MOUSAVI R, JOHAR M, MOOKERJEE V S. The Voice of the Customer: Managing Customer Care in Twitter. ***Information Systems Research***, 2020, 31(2):340-360.

[26] 赵卫宏,熊小明. 网络零售服务质量的测量与管理——基于中国情境. 管理评论, 2015, 27(12):120-130.

[27] CAIGNY A D, COUSSEMENT K, BOCK K W, LESSMANN S. Incorporating textual information in customer churn prediction models based on a convolutional neural network. ***International Journal of Forecasting***, 2019, 36, 1563-1578.

[28] VO N N, LIU S, LI X, XU G. Leveraging unstructured call log data for customer churn prediction. ***Knowledge Based System,*** 2021, 212, 106586.

[29] Blei, Ng & Jordan. Latent Dirichlet Allocation. ***Journal of Machine Learning Research***, 2003, 3(4-5):993-1022.

[30] SLOF D, FRASINCAR F, MATSIIAKO V. A competing risks model based on latent dirichlet allocation for predicting churn reasons. ***Decision Support Systems***, 2021, 146, 113541.

[31] KWON H, KIM H H, AN J, LEE J H, PARK Y R. Lifelog Data-Based Prediction Model of Digital Health Care App Customer Churn: Retrospective Observational Study. ***Journal of medical Internet research***, 2021, 23(1), e22184.

[32] Atelsek, F J, Schutz W C. Firo: A Three-Dimensional Theory of Interpersonal Behavior. 1959.

[33] 卫海英, 杨国亮.企业-顾客互动对品牌信任的影响分析——基于危机预防的视角[J].财贸经济,2011,(4):79-84.

[34] Mcmijlan S J. The researchers and the concept: moving beyond a blind examination of interactivity[J]. ***Journal of Interactive Advertising***，2005，5(1).

[35] Yen Y S. The Impact of Perceived Value on Continued usage Intention in Social Networking Sites[J]. ***2011 2nd International Conference on Networking and Information Technology，IPCSIT***，2011，(17):217 -223.

[36] 汪旭晖, 张其林. 用户生成内容质量对多渠道零售商品牌权 益的影响. 管理科学, 2015, 28(4): 71−85.

[37] 韩雨彤，周季蕾，任菲.动态视角下实时评论内容对直播电商商品销量的影响[J].管理科学,2022,35(1):17-28.

[38] 孟庆斌, 黄清华, 赵大旋, 鲁冰. (2019). 互联网沟通与股价崩盘风险. 经济理论与经济管理, 1000-596X, (11), p. 50.

[39] 卞世博，陈曜，汪训孝. 高质量的互动可以提高股票市场定价效率吗？基于“上证 e 互动”的研究. 经济学(季刊), 2022, 749-772.

[40] Parasuraman, A., Zeithaml, V.A. and Berry, L.L. SERVQUAL: A Multiple-Item Scale for Measuring Consumer Perceptions of Service Quality. ***Journal of Retailing,*** 1988, 64, 12-40.

[41] 刘俊清. B2C 环境下线上互动、感知价值与消费者重购意愿的关系研究[J]. 内蒙古财经大学学报,2018,16(2):50-55.

[42] Leech G.Prineiples of Pragmatics[M].London:Longman, 1983.

[43] 黄裕乐.移动互联网背景下的淘宝售前客服沟通效能分析. 现代营销(营销版). 2018. No.308(08), 105.

[44] 刘琼.移动电子商务互动营销及应用模式研究. 电子商务, 2018, (9), 37-38.

[45] Vlasic, Goran & Kesic, Tanja. Analysis of Consumers' Attitudes toward Interactivity and Relationship Personalization as Contemporary Developments in Interactive Marketing Communication. ***Journal of Marketing Communications.*** 2007. 13. 109-129. 10.1080/13527260601070417.

[46] 孙皓, 任俊生, 宋平平. (2011). 营销沟通方式与顾客忠诚:一个研究综述. 现代管理科学, 2011(1):27-29.

[47] 张爱萍，杨东帆.(2022).互动营销视角下微博广告的信息特征对品牌价值的影响. 经营与管理, (07),46-51.

[48] Wang, F., Liu, J., and Wang, H. Sequential Text-Term Selection in Vector Space Models. ***Journal of Business and Economic Statistics.*** 2021, 39(1):82-97.

[49] Yang, Y. and Wang F. Author Topic Model for Co-occurring Normal Documents and Short Texts to Explore Individual User Preferences. ***Information Sciences***. 2021, 570: 185-199.

[50] CHAN, ELAINE and SENGUPTA, JAIDEEP. Observing flattery: a social comparison perspective. ***Journal of Consumer Research.*** 2013

[51] Xiaodong, Li, Xinshuai, Guo, Chuang, and Wang, et al. Do buyers express their true assessment? antecedents and consequences of customer praise feedback behaviour on taobao. ***Internet Research.*** 2016, 26(5), 1112-1133.

[52] Danziger, R. The pragmatics of flattery: the strategic use of solidarity- oriented actions. ***Journal of Pragmatics.*** 2020, 170, 413-425.

附 录

附表 1：模型 1 的估计结果

变量	估计值	变量	估计值
常数项	-5.671***	所在地_北京市	0.391***
客户历史续费次数	0.231***	所在地_深圳市	0.231**
第一次付费距今月数	-0.031***	上次续费所购课时_30 至 50	0.151*
剩余课时数	0.331***	上次续费所购课时_50	-0.191***
该月月初过去 45 天完课率	1.001***	孩子在业务内的年级_2.0	-0.181**
该月月初过去 45 天缺课率	-1.071***	孩子在业务内的年级_3.0	-0.291***
该月月初过去 45 天课时消耗数	0.021*	孩子在业务内的年级_4.0	-0.311***
是否有优惠券	1.001***	是否有代币	-0.311***
是否在班	1.141***	剩余课时数_平方	-0.051***
课后练习完成率	0.141**	剩余课时数_立方	0.001**
测验参与率	0.121*	剩余课时数_四次方	0.001*
孩子综合表现水平值	0.001**	有无代币与课时数交叉项_平方	0.011**
测验题目正确率	0.491*	有无代币与课时数交叉项_立方	0.001**
平均课堂答题次数	0.031**	家长使用 APP 查看学情报告次数	0.281*
平均课堂答题正确数	1.421***	聊天记录平均情感得分	2.281***
朋友圈海报分享完成率	-0.371***	聊天记录平均情感得分_平方	-2.941***
距上次家长发消息天数	0.001***	聊天记录条数	0.001***
客服消息占总消息比例	-0.291**		0.001

注: ***为 0.001 显著性水平，**为 0.010 显著性水平，*为 0.050 显著性水平

附表 2：模型 2 的估计结果

变量	估计值
常数项	-3.341***
主题变量 1：客服礼貌用语	6.881***
主题变量 2：促销期：预设消息	-3.391***
主题变量 3：日常：预设消息	-6.261***
主题变量 5：课程协调	5.981***
主题变量 6：促销期：续费	0.961**
主题变量 7：日常：续费	2.421***
主题变量 8：学生测评	5.601***
主题变量 12：教学内容反馈	0.181***

注: ***为 0.001 显著性水平，**为 0.010 显著性水平，*为 0.050 显著性水平