北京大学

题目：基于 RMF 时序预测在线零售商的客户盈利能力

课程名称：	量化营销模型
授课老师：	刘宏举
成员姓名：	彭潘杰、沈逸然
完成日期：	2022.6.15

二〇二二年六月

摘要

本文对客户盈利能力随时间的动态预测进行了研究分析。通过收集一家英国零售店的真实交易数据集，我们使用近期购买行为、购买的总体频率以及花费金额（RFM）模型来衡量客户的盈利能力，并相应地生成该企业每个客户的月度 RFM 时间序列。在每个时间点，通过使用 k 均值聚类并比较不同类别的盈利能力，将顾客的 RFM 划分为高、中或低组，并统计随着窗口期的变化不同盈利能力的客户数量，发现随着考虑的时间周期越来越长，不同盈利能力的客户占比基本保持稳定。除此之外，聚类分析还通过给定不同窗口期不同客户的标签给出了每一个客户盈利能力的动态变化过程，为下一步使用时间序列的机器学习模型预测客户未来的盈利能力提供了数据。

为了进一步对客户进行有针对性的营销，我们训练了循环神经网络模型并发现了该机器学习模型预测零售店顾客盈利能力的高准确性。商家可以利用这一预测的数据进行针对性的营销。对于即将流失的客户使用促销等手段防止其流失，对于未来盈利能力会上涨的客户可以使用推销的方式进一步增强其盈利能力等等。

关键词：动态预测消费者盈利能力、RFM 顾客分类、Kmeans 聚类、循环神经网络

第一章 研究目标

随着时间的推移，动态预测消费者的盈利能力在当今以客户为中心的商业模式中起着至关重要的作用。企业通常会在给定的时间点评估客户的盈利能力，以便针对某些特定客户进行营销。然而实际上，客户的购买习惯和偏好是多种多样的，并且可能会随着时间的推移而产生变化。因此，企业希望根据每个客户的历史购买和盈利能力推算他/她的盈利能力会如何随着时间的推移而发展，而这个问题的答案也将直接影响企业的营销策略和资源分配。一般来说，做出这样的预测需要考虑几个主要方面，包括：

（1）指定衡量消费者盈利能力的指标。

（2）给定消费者的购买历史记录和指定的盈利能力指标，构建适当的动态模型来描述客户盈利能力的动态，并进一步预测消费者盈利能力。

在商业背景下，有许多评估客户行为和客户盈利能力的指标，例如近期购买行为、购买的总体频率以及花费金额（RFM）模型、客户生命周期价值（CLV）模型和帕累托（Pareto）模型（Januszewski，2011）。每个指标都提供了一个独特的视角，不同指标之间也存在一些显式或隐式关系。某个指标的选择通常取决于企业业务所关注的内容，有时可能会应用多个指标。

本文选取的数据来自一家英国在线零售商的一组交易记录。我们选取 RFM 指标作为客户盈利能力的衡量标准，每个客户都被分配了一个独特的 RFM 分数以反映客户的购买行为和盈利能力。通过在给定的时间段内将客户按 RFM 值聚类并分配统一值，我们可以挖掘客户盈利能力的潜在特征。同时，针对不同时间段的客户盈利能力的变化可以构建具有特定时间段标签的客户盈利能力值的时间序列，并用于训练和测试神经网络模型。

在本文中，我们通过训练可循环神经网络（RNN）来捕捉客户盈利能力，从而探索动态预测客户盈利能力的问题以得到更多有关客户购买行为的洞见，提升商家的营销能力。

第二章 数据描述

（一）数据来源

本文使用的数据集来自一家英国注册的在线零售商（Chen et al., 2012）。数据集中有 11 个变量，变量的属性以及具体含义如表 1 所示，它包含了 2009 年到 2011 年发生的所有交易。在 2009 年 12 月到 2011 年 12 月这段期间这段时间总共产生了 53628 笔有效交易，包括 5943 名消费者购买的共计 5305 件产品。

表 1 变量定义

变量名称	变量类别	变量描述
Invoice	名义	唯一分配给每笔交易的 6 位整数。如果此代码以字母“c”开头，则表示取消。
Stock Code	名义	一个 5 位整数，唯一分配给每个不同的产品。
Description	名义	产品名称
Quantity	数值	每笔交易中每种产品（项目）的数量
InvoiceDate	数值	生成交易的日期和时间。
Price	数值	每单位产品价格（英镑）
Customer ID	名义	唯一分配给每个客户的 5 位整数。
Country	名义	客户居住的国家/地区的名称。

（二）RFM 变量描述

RFM 分别指的是每名顾客距离上次购买的时间长度，一段时间内后买的频率以及消费的总金额数，该指标可以反应顾客的活跃度、忠实度与购买力。本文创建了一个时间序列的数据集。由于所考虑的问题是时间序列的预测问题，目标数据集应具有固定大小，包含在整个分析时间段至少与零售商进行过一次购买的客户的购买记录。计算 RFM 值的截止时间点设置为 2009 年 12 月至 2011 年 11 月的每个日历月的月底。随后，每个客户最多有 24 组 RFM 值相关联，每组对应一个特定的时间段，即 2009 年 12 月、2009 年 12 月至 2010 年 1 月、2009 年 12 月至 2010 年 2 月、……以及 2009 年 12 月至 2011 年 11 月。最后一期 RFM 三个指标的统计结果见表 2 和图 1：

表 2 RFM 描述性统计

	Recency	Frequency	Monetary
count	5914	5914	5914
mean	200.38671	133.902435	2705.63826
std	208.918388	343.952743	13442.5124
min	0	1	-25111.09
25%	23	20	319.1675
50%	100.5	52	817.735
75%	373	138.75	2120.4625
max	729	12436	558895.07

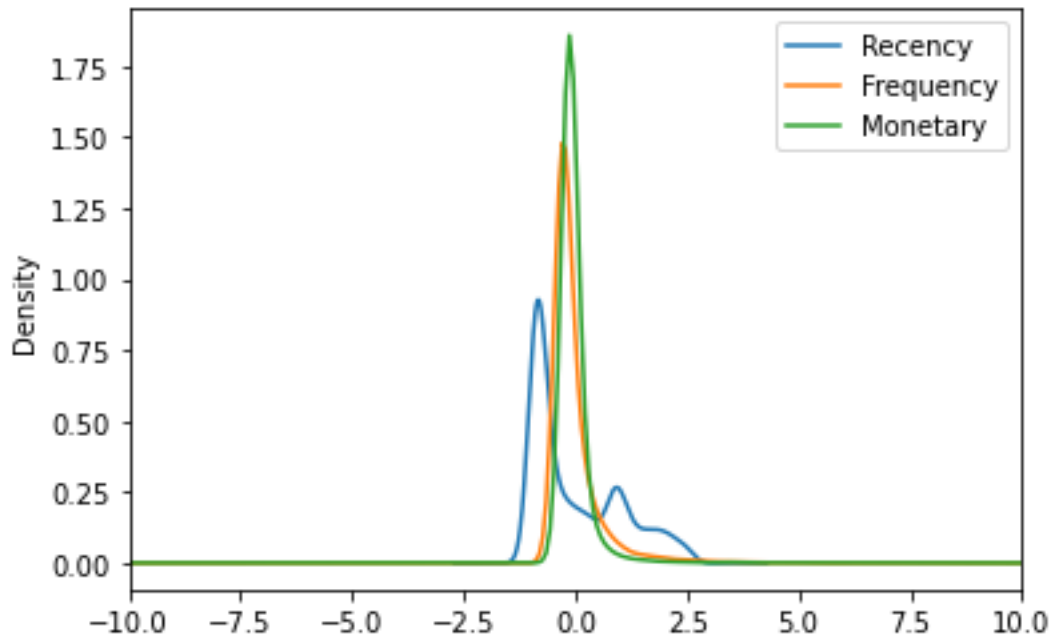

图 1 标准化后的 RFM 数据的概率密度函数图

从表 2 和图 1 中我们可以看出大部分数据在标准化（原数据减均值除以标准差）之后分布在正负三倍标准差之间，但仍然存在小部分离群值，为防止其对 kmeans 均值聚类结果产生过大影响，将大于 3 倍标准差的值全部赋值为 3，小于三倍标准差的值全部赋值为 -3。并对标准化后的 RFM 数据进行 Kmeans 聚类。

第三章 数据分析

（一）使用 RFM 数据进行聚类分析

根据每个时间段结束时的 RFM 值，相关客户被分为五个组（集群）k-means 聚类算法。每个集群都包含一组具有相似 RFM 值的客户。然后，这些 RFM 值被汇总在一起，以确定同一集群中所有客户的唯一单值 RFM 分数。

表 3 给出了在任何给定时间段结束时用于确定同一集群中客户的 RFM 分数的规则。

表 3 聚类分数确定规则

聚类标签	RFM 分数	客户盈利能力
1	1	高
2,3,4	2	中
5	3	低

在整个分析时间段内，每个客户都与最多 24 个 RFM 分数相关联，这些分数形成了单个客户的基于 RFM 分数的盈利时间序列。图 2 显示了实验中发现的几个典型的 RFM 分数时间序列。很明显，客户的盈利能力通常会随着时间的推移而变化——在某些情况下非常明显，并且分数的动态演变是多种多样的。然而，另一方面，如表 4 和图 5 所示，不同盈利组的客户数量随着时间的推移变得稳定。

表 4 2010.1-2011.12 不同盈利能力客户数量

	1	**2**	**3**
2010/1/1	262	675	108
2010/2/1	529	756	154
2010/3/1	513	1117	172
2010/4/1	383	1620	235
2010/5/1	350	1916	263
2010/6/1	482	2022	279
2010/7/1	485	2235	332
2010/8/1	502	2382	351
2010/9/1	582	2458	353
2010/10/1	662	2608	365
2010/11/1	587	2967	460
2010/12/1	578	3297	461
2011/1/1	607	3301	505
2011/2/1	749	3302	433
2011/3/1	913	3311	383
2011/4/1	895	3493	397
2011/5/1	835	3610	445
2011/6/1	823	3708	467
2011/7/1	810	3835	461
2011/8/1	794	3923	491

2011/9/1	792	4046	476
2011/10/1	815	4222	464
2011/11/1	811	4395	516
2011/12/1	790	4585	539

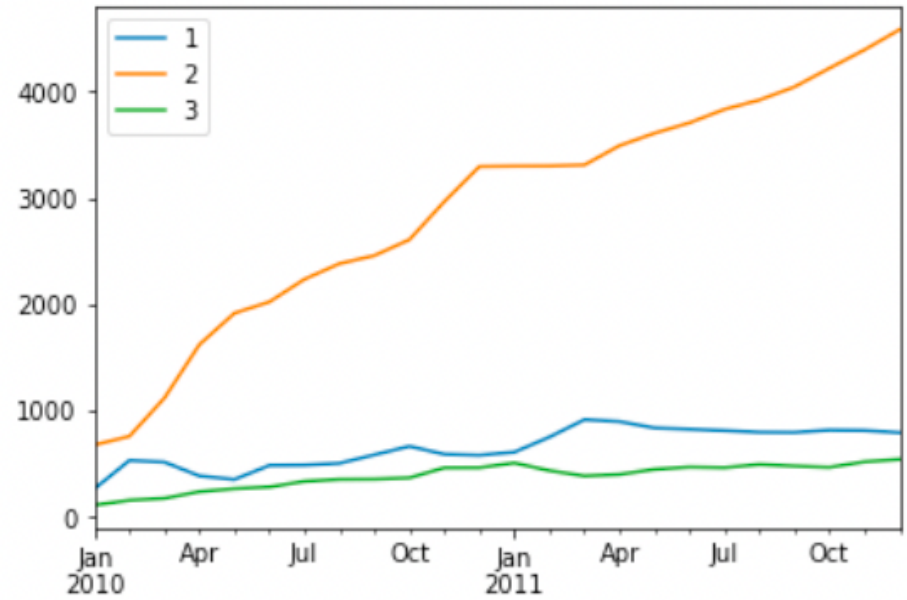

图 2 2010.1-2011.12 不同盈利能力客户数量变化

（二）使用循环神经网络模型进行训练和预测

首先，通过数据预处理剔除数据量不够的样本，即顾客在每个分析时间段至少与零售商进行过一交易。前文提到过，计算 RFM 值的截止时间点设置为 2009 年 12 月至 2011 年 11 月的每个日历月的月底。因此，在处理过后的数据中，每个客户都有 24 组 RFM 值相关联。

为了解决动态预测客户盈利能力的问题更好的了解顾客的购买行为并进行更精准的营销活动，因为循环神经网络（RNN）相比于普通的神经网络具有更好的学习时间序列特性的能力，因此，本报告使用循环神经网络（RNN）模型的长短期记忆网络（LSTM）（Manaswi，2018）使用客户基于 RFM 分数的盈利时间序列数据进行将来的预测。

在每个时间戳处，先前在时间窗口中的数据首先会被前馈到全连接神经网络（FC）中，FC 的输出会接着作为 LSTM 单元的输入。LSTM 单元的输出会进一步馈送到另一个全连接神经网络（FC）并生成当前时间戳的预测。在本报告中，我们用 Pytorch 实现了 LSTM，同时，我们将最后一个时间戳的数据作为测试集，并把之前时间段的数据作为训练集，得到如下图的训练曲线：

图 3 LSTM 的训练损失

在测试集上验证模型效果，得到的训练的准确度是：0.9665。混淆矩阵如下图所示：

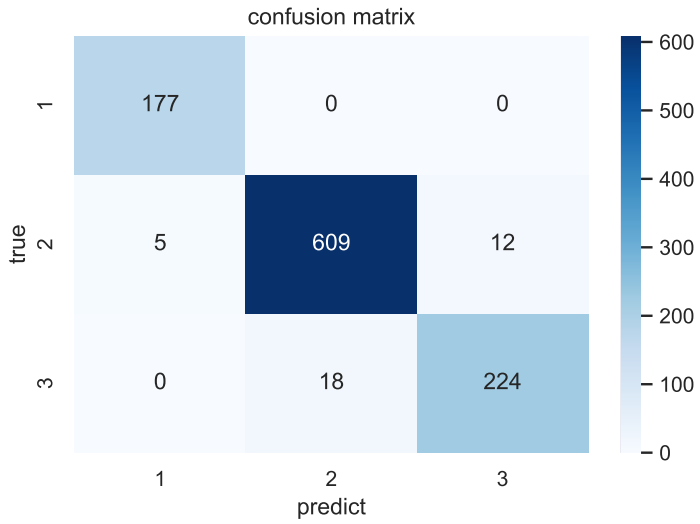

图 4 测试集上预测结果的混淆矩阵

由此可见，使用 LSTM 进行预测可以达到很高的准确率，因此使用循环神经网络可以精准的预测客户盈利能力。

第四章 结论

在本篇报告中，我们提出了一种预测客户盈利能力的新方法。 该方法的主要思想是选取构建一组基于 RFM 分数的时间序列来表示客户盈利能力的动态过程。本文使用 k 均值聚类并比较不同类别的盈利能力，将顾客的 RFM 划分为高、中或低组，并统计随着窗口期的变化不同盈利能力的客户数量，发现随着考虑的时间周期越来越长，不同盈利能力的客户占比基本保持稳定。除此之外，聚类分析还通过给定不同窗口期不同客户的标签给出了每一个客户盈利能力的时间序列数据。这些结果表明，基于 RFM 分数的时间序列反映了与客户购买偏好和习惯相关的一些内在特征。此外，结果也表明，基于 RFM 分数的客户盈利能力是可预测的，此处使用了 RNN 模型进行训练，并得到了较高的预测精度。然而，该神经网络是一个黑盒过程，并不能观察之间的学习过程，导致输出的结果难以解释并进一步影响结果的可行度和可接受程度。

根据我们模型的预测结果，管理者可以辨别不同客户的特征，并进行针对性的营销手段。比如商家可以利用这一预测的数据进行针对性的营销。对于即将流失的客户使用促销等手段防止其流失，对于未来盈利能力会上涨的客户可以使用推销的方式进一步增强其盈利能力等等。

我们认为，这篇报告所提出的方法具有前景，特别是在以下领域值得进一步的研究：

- 调查不同盈利组的客户购买了哪些产品，以探索预测客户在下一个时间段可能购买哪些产品的可能性。
- 用于预测基于 RFM 分数的时间序列的替代模型，例如动态贝叶斯网络。

参考文献

Ale, L., Zhang, N., Wu, H., Chen, D., & Han, T. (2019). Online Proactive Caching in Mobile Edge Computing Using Bidirectional Deep Recurrent Neural Netwoek. *IEEE Internet of Things Journal*, (6), 5520-5530.

Chen, D., Sain, S.K., & Guo, K. (2012). Data Mining for the Online Retail Industry: A Case Study of RFM Model-Based Customer Segmentation Using Data Mining. *Journal of Database Marketing and Customer Strategy Management*, (19), 197-208.

Chen, D., Guo, K., & Ubakanma, G. (2015). Predicting Customer Profitability Over Time Based on RFM Time Series. *International Journal of Business Forecasting and Marketing Intelligence*, (2), 1-18.

Chen, D., Guo, K., & Li, B. (2019). Predicting Customer Profitability Dynamically Over Time: An Experimental Comparative Study, Based on RFM Time Series. *24TH Iberoamerican Congress on Pattern Recognition, Cuba,* 28-31.

Januszewski, F. (2011). Possible Applications of Instruments of Measurement of the Customer Value in the operations of Logistcis Companies. *Scientific Journal of Logistics*, (7), 17-25.

Manaswi, N.K. (2018). RNN and LSTM. *Deep Learning with Application Using Python*, Apress.

附录

传统的深度神经网络不能保存从以前的时段里学习到的信息，这也是神经网络（NN）解决序列问题存在的问题之一。循环神经网络（RNN）则可以通过循环将学习到的信息从一个步骤传递到下一个步骤。长短期记忆网络（LSTM）是一种特殊的 RNN，主要是为了解决长序列训练过程中的梯度消失和梯度爆炸问题，可以在更长的序列中有更好的表现。