客户发展关系的马尔可夫过程模型及其应用

一、问题背景

1.1 客户关系管理的现状分析

随着市场竞争的日益加剧，客户已经越来越成熟，市场已经完成了由卖方市场向买方市场的转型。企业越来越强烈地感觉到客户资源将是企业竞争中至关重要的资源，客户才是企业生存发展的源泉，于是，把客户作为企业的一项资产来经营并研究如何获得客户并且是有价值的客户，这就必须要研究客户关系。
20 世纪 90 年代以来，客户关系管理(Customer Relationship Management，CRM) 已经成为企业营销策略研究和营销系统应用的持续热点。为客户创造价值是客户关系管理的核心和出发点。客户关系管理的目标并不是最大化单笔交易的价值，而是构筑与客户之间长期持久的关系，培养忠诚的客户。企业通过为客户创造价值，增加客户的满意度，从而增加客户对企业的忠诚度，增加商机，进而实现双方价值的最大化。
本文针对客户关系发展规律，用 Markov 链建立客户关系发展的一般数学模型，并从 Markov 链的转移概率矩阵开始进行研究，对 CRM 中客户关系发展模型的建立进行了深入分析。并且，本文的研究结果可为企业的客户关系管理提供参考，有一定研究意义。

1.2 RFM 模型简介

RFM 模型是衡量客户价值和客户创利能力的重要工具和手段。在众多的客户关系管理(CRM)的分析模式中，RFM 模型是被广泛提到的。该模型通过一个客户的最近一次消费、消费频率、消费金额 3 项指标来描述该客户的价值状况。在 RFM 模式中，R(Recency)表示客户购买的时间有多远，F(Frequency)表示客户在时间内购买的次数，M(Monetary)表示客户在时间内购买的金额。本文将考虑这 R 指标作为对客户状态的分类依据。

1.3 马尔可夫决策分析

客户关系管理需要衡量客户与公司的关系在多个时期的连续变化，由于马尔可夫模型处理时间序列的信息能力较强，马尔可夫决策过程作为一个完善的数学分支学科在解决问题上已经成熟。在客户购买行为中，客户在一个给定的时期决定是否购买及购买的数量，在购买决策只是当前客户状态和公司策略的函数的假设下，形成客户购买值序列。由于客户状态的发展具有无后效性，即客户关系下一阶段处于哪个阶段，只与它现在处于哪个阶段有关，而与之前的状态无关，这显然具备了马尔可夫分析方法的使用条件，故可以将马尔可

夫决策过程应用动态客户关系管理。客户关系的发展是有阶段性的，本文就客户关系的阶段性发展进行定量研究并运用离散的马尔可夫链相关理论建立模型，预测客户的未来行为。

二、建立模型

2.1 马尔可夫链简介

设离散时间参数随机过程$\{X_n, n \geq 0\}$的状态空间也是离散的，记为

$$E = \{0,1,2,\ldots\}, \forall i_0, i_1, \cdots, i_{n-1}, i, j \in E, \forall n \geq 0,$$都有：

$$P(X_{n+1} = j | X_0 = i_0, \cdots, X_{n-1} = i_{n-1}, X_n = i) = P(X_{n+1} = j | X_n = i)$$

则称$\{X_n, n \geq 0\}$为 Markov 链。

下面我们介绍最常用的一种 Markov 链，时齐 Markov 链：

$\forall i, j \in E, P(X_{n+1} = j | X_n = i) = p_{ij}(n)$ 为 n 时刻的一步转移概率。如果$p_{ij}(n) \equiv p_{ij}$,而与 n 无关，则称$\{X_n, n \geq 0\}$为齐次 Markov 链。本文只讨论齐次马尔可夫链。

记$P = (p_{ij})_{i,j \in E} = \begin{bmatrix} p_{00} & \cdots & p_{0n} \\ \vdots & \ddots & \vdots \\ p_{n0} & \cdots & p_{nn} \end{bmatrix}$, 则称 P 为马尔可夫链$\{X_n, n \geq 0\}$的一步转移概率矩阵。

状态转移概率矩阵是一个 n 阶方阵,它具有下述性质：

$(1) p_{ij} \geq 0 \ (i, j = 1,2,\cdots, n)$，即每个元素均是非负的。

$(2) \sum_{j=1}^{n} p_{ij} = 1 \ (i, j = 1,2,\cdots, n)$，即矩阵每行的元素和等于 l。

m 步转移矩阵为：

$$P_{ij}^{(m)} = P(X_{n+m} = j | X_n = i) = P(X_m = j | X_0 = i)$$

表示当前时刻过程在状态 i，经过 m 步后到达了状态 j 的概率。

m 步转移概率矩阵的两个性质为：

1．$P_{ij}^{(0)} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$

2．C-K 方程：

$$P_{ij}^{(m+n)} = \sum_{k \in E} P_{ik}^{(m)} P_{kj}^{(n)} \quad 或 \quad P^{(m+n)} = P^{(m)} \times P^{(n)}$$

记$\pi_i(n) = P(X_n = i), \pi(n) = (\pi_1(n), \pi_2(n), \cdots, \pi_i(n), \cdots),$

$\pi(n)$表示n时刻X_n的概率分布向量。称$\{\pi_i(0), i \in E\}$为Markov链的初始分布。容易验证：

$$\pi(n+1) = \pi(n)P$$
$$\pi(n) = \pi(0)P^n$$

因此 ,时齐马氏链的概率分布完全由初始分布$\pi(0)$和概率矩阵P决定。

2.2 客户状态的确定

我们要建立的客户关系发展的马尔可夫过程模型是基于相似客户群体的,即具有许多相似的特征(如需求数量等)的客户, 因此首先要确定不同客户所处的状态。考虑到 RFM 因素，由于消费金额对客户的关系没有那么显著，并且消费金额为连续值难以将其分类，且 F 指标的值可以用 R 所表示，所以本文以最近一次消费 R 这项指标进行状态的确定。然后利用这个指标 R 确立 6 了个最终的客户状态。
对于最近一次消费有多远的近度 R，下面为我们确定的 6 种状态：
0: 从未买过公司的产品
1: 当前时间段买了公司的产品
2: （当前时段没有买）在过去的一个时间段内买了公司的产品
3. （当前时段和过去的一个时间段内没有买）过去两个时间段内买了公司的产品
4. （当前时段过去的两个时间段内没有买）过去三个时间段内买了公司的产品
5: （当前时段与过去的一个、两个或三个时间段内没有买）在过去的四个时间段之前（包括第四个）买了公司的产品(n>=4)
状态空间 R:{0,1,2,3,4,5}

最后，我们设定状态 5 为一个客户的流失状态，即三年以上未买过公司产品的客户，我们认为处于流失状态的客户不会再回来与该企业发展客户关系，即如果一个客户三年内都没有购买公司产品，那么我们认为这个客户永远不会回来再买公司的产品。所以，我们一共定义了 6 个用户状态，即状态空间为 E={0,1,2,…,5}.
最终指标确定出的客户状态如下所示：
```
{0: 0, 1: 1, 2: 2, 3: 3, 4: 4, 5: 5}
```
我们说明本文中马尔可夫链建模的基本原理并用一个例子来说明具体模型：
我们研究的是一个状态空间 E 离散，时间参数集 T 也离散的一个离散随机序列 Markov 链。在这里我们设定 Markov 链的时间域为 T={2005,2006,…,2015}，为方便起见，我们重新写为 T={0,1,2,…,10}，共 11 个时间状态。即以一年为一个单位时期，从前一个状态转移到下一个状态是过去一年转移到当前的年。请考虑以下情况：直销公司正试图说服艾米作为客户。如果成功，直销公司希望在艾米最初的购买和每次成功的购买中获得对公司的利润。艾米可能在两次购买之间隔了好几个时期，期间长度相等为一年。该公司认为，艾米在任何时期结束时购买的可能性仅是艾米的近度（即艾米上次购买以来的时期数）的函数。如果艾米在上期购买了，则艾米在当前期间的近度为 2。令艾米在任何时期末购买的概率为P_r，其中 r 是艾米的近度。为了给出清楚说明，假设艾米无论何时到达 r=5，公司将会把艾米归于流失客户状态。因此，在任何时期结束时，公司与艾米的关系有六种可能的状态，潜在客户状态 0(r=0)，当前期购买 r=1,分别对应于近度为 2 至 4 的三种状态

和第五种状态，即“非客户”或“前客户”，我们将其标记为 r=5。公司与艾米关系的一个关键特征是，这种关系的未来前景仅取决于关系的当前状态（由艾米的近度），而不取决于艾米的通过哪条特定路径达到了她目前的状态。此属性称为 Markov 属性。马尔可夫属性是随机系统成为马尔可夫链的必要条件。图 1 展示了艾米与此公司的关系转移图。

2 .3 客户转移概率的确定

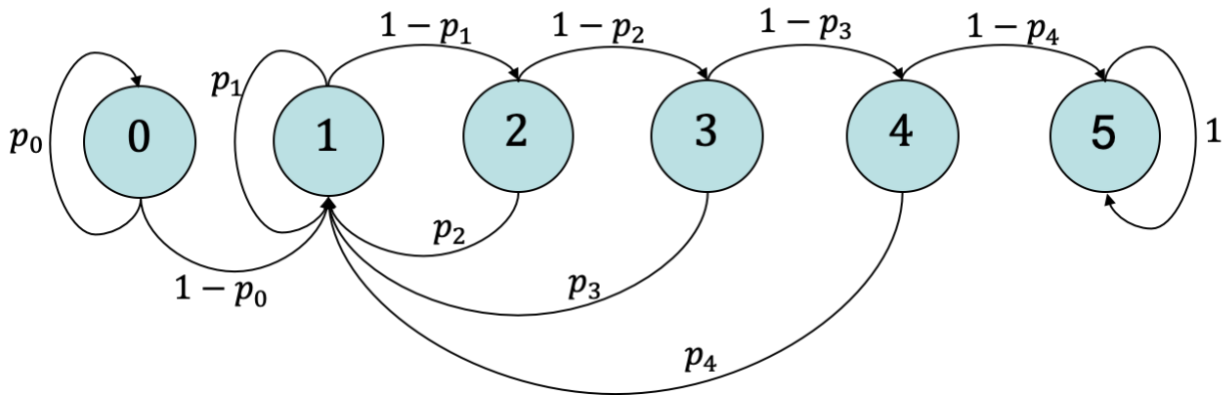

图 1：该公司与艾米的关系的跳转图表示

在某时刻t_0，当已知客户所处的状态 i 时，客户在时刻 $t(>t_0)$ 时所处的状态与客户在时刻之前所处的状态无关，因此客户关系的发展过程具有无后效性，故可用马尔可夫过程表示。一个客户在某一时刻 t 只能处于上述状态集中的某一状态 i，在下一时刻 t +1 客户将以转移概率P_{ij}转移到状态 j。我们认为，除客户处于流失状态外，客户关系发展层次的上升是逐层递进的，不会越过一个层次而进入更高层次。且客户一旦购买过该公司产品就不会再返回到状态 0(潜在客户状态)。

我们可以用一个6×6的转移概率矩阵来总结图 1 中以跳转图方式显示的转移概率。该矩阵的最后一行反映了以下假设：如果艾米在任何将来的时间段到达近度 5，那么她在下一个和所有未来期间都将保持在近度 5。用马尔可夫链的语言，r=5 或流失客户状态是一个吸收态。艾米一旦进入该状态后，她将一直保持在该状态。

各状态间的一步转移概率矩阵 P 为：

$$P^{(1)}=\begin{array}{c|cccccc} & 0 & 1 & 2 & 3 & 4 & 5 \\ \hline 0 & p_0 & 1-p_0 & 0 & 0 & 0 & 0 \\ 1 & 0 & p_1 & 1-p_1 & 0 & 0 & 0 \\ 2 & 0 & p_2 & 0 & 1-p_2 & 0 & 0 \\ 3 & 0 & p_3 & 0 & 0 & 1-p_3 & 0 \\ 4 & 0 & p_4 & 0 & 0 & 0 & 1-p_4 \\ 5 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}$$

矩阵 P 是一阶转移概率矩阵。m 步转移概率矩阵定义为恰好在 m 个周期 内从一种状态迁移到另一种状态的概率矩阵。由 m 步转移概率矩阵的性质 2（C-K 方程）我们知道 m 步转移矩阵只是 m 个一步转移矩阵的矩阵乘积。从而 $P^{(m)} = P^m$

2 .4 客户关系发展的马尔可夫模型

下面给出通过对实证分析案例进行建模并且计算得到的一步转移概率矩阵，在这里本文先分析此过程的性质，在第三章详细说明实证分析的过程。我们研究的模型是一个有限状态马氏链, $E = \{0,1,2,\dots,5\}$, 客户状态到达 5 就吸收不再转移, $\{X_n = $ 时刻 n 客户所处的状态$\}$

P=	0	1	2	3	4	5
0	0.8	0.2	0.0	0.0	0.0	0.00
1	0.0	0.5	0.5	0.0	0.0	0.00
2	0.0	0.2	0.0	0.8	0.0	0.00
3	0.0	0.1	0.0	0.0	0.9	0.00
4	0.0	0.05	0.0	0.0	0.0	0.95
5	0.0	0.0	0.0	0.0	0.0	1.00

经实际数据计算得到马氏链的初始状态分布为

$$\pi(0) = \{\pi_i(0), i \in E\} = (p_0, p_1, p_2, p_3, p_4) = (0.9335, 0.0665, 0, 0, 0)$$

又有$\pi(n) = \pi(0)P^n$，我们可以用这个模型来预测未来年度的顾客状态。

- **可达与互通**

为分类做准备，我们先研究各个状态之间的可达和互通关系。

$$0 \to 1, 1 \not\to 0$$

$$1 \leftrightarrow 2, 1 \leftrightarrow 3, 1 \leftrightarrow 4$$

由可达关系与互通关系都具有传递性，有$0 \to 2, 0 \to 3, 0 \to 4, 0 \to 5$

首达时间：

$$\forall i, j \in E, T_{ij} = \begin{cases} \min\{n: X_0 = i, X_n = j, n \geq 1\} \\ \infty, \text{若以上集合为空} \end{cases}$$

那么我们就知道$T_{10} = \infty, T_{20} = \infty, T_{30} = \infty, T_{40} = \infty, T_{50} = \infty$。

- **首达概率**

首达概率$f_{ij}^{(n)}$表示从状态 i 出发,经过 n 步首次到达状态 j 的概率，称为首达概率。

$$f_{ij}^{(n)} = P\left(T_{ij} = n \middle| X_0 = i\right) = P(X_n = j, X_k \neq j, 1 \leq k < n | X_0 = i)$$

$$f_{00}^{(1)} = p_0 = 0.8, f_{00}^{(2)} = 0, f_{00}^{(3)} = 0, \cdots, f_{00}^{(n)} = 0;$$

$$f_{11}^{(1)} = p_1 = 0.5, f_{11}^{(2)} = 0.5 \times 0.2 = 0.1, f_{11}^{(3)} = 0.5 \times 0.8 \times 0.1 = 0.04,$$

$$f_{11}^{(4)} = 0.5 \times 0.8 \times 0.9 \times 0.05 = 0.018, \ \ f_{11}^{(5)} = 0, \cdots, f_{00}^{(n)} = 0;$$

$$f_{22}^{(1)} = 0, f_{22}^{(2)} = p_2 \times (1 - p_1) = 0.2 \times 0.5 = 0.1,$$

$$f_{22}^{(3)} = p_2 \times p_1 \times (1 - p_1) = 0.2 \times 0.5 \times 0.5 = 0.05,$$

$$f_{22}^{(4)} = 0.2 \times (0.5)^2 \times 0.5, \cdots f_{22}^{(n)} = 0.1 \times (0.5)^{n-2}, \cdots$$

- **周期性**

设$\{X_n, n \geq 0\}$为齐次马氏链，其状态空间为 E. $\forall i \in E$,若集合$D = \{n: p_{ii}^{(n)} > 0, n \geq 1\}$非空，令 d(i)为该数集的最大公约数，称其为状态 i 的周期；若上述集合为空集，令$d(i) = \infty.$
若$d(i) > 1$，称状态 i 为有周期的；
若$d(i) = 1$，称状态 i 为非周期的；
本模型中，显然 d(0)=1,d(1)=1,d(5)=1，所以状态 0,1 和 5 为非周期的。
状态 2:
$D = \{2,3,4,5, \ldots\} \Longrightarrow d(2) = 1$
状态 3:
$D = \{3,4,5,6, \ldots\} \Longrightarrow d(3) = 1$
状态 4:
$D = \{4,5,6,7, \ldots\} \Longrightarrow d(4) = 1$
所以状态 2,3 和 4 也为非周期的。
其实，我们由下面的性质可以快速的得到这个结论：

周期的性质：若$i \leftrightarrow j$,则$d(i) = d(j).$

由于状态 1 和状态 2，3，4 均互通，则它们的周期一致，均为 1。

- **常返性：**

$f_{ij} = \sum_{n=1}^{\infty} f_{ij}^{(n)}$,表示过程由状态 i 出发，经过有限步终于到达状态 j 的概率。

则我们可以用如下的方法判断状态的常返性：

- 若$f_{ii} = 1$，则称状态 i 为常返的。
- 若$f_{ii} < 1$，则称状态 i 为非常返的，或瞬时的。

状态 0:

$f_{00} = \sum_{n=1}^{\infty} f_{00}^{(n)} = f_{00}^{(1)} = 0.8 < 1$, 所以状态 0 非常返；

状态 1:

$f_{11} = \sum_{n=1}^{\infty} f_{00}^{(n)} = 0.5 + 0.1 + 0.04 + 0.018 = 0.658 < 1$，所以状态 1 非常返；

状态 2:

$f_{22} = \sum_{n=1}^{\infty} f_{00}^{(n)} = 0 + 0.1 \times \sum_{i=2}^{\infty} (\frac{1}{2})^{n-2} = 0.1 \times \frac{1}{1-\frac{1}{2}} = 0.2 < 1$, 所以状态 2 非常返；

状态 3: 容易验证，状态 3 非常返；
状态 4: 容易验证，状态 4 非常返；

状态 5: $f_{55} = f_{55}^{(1)} = 1$，所以状态 5 是常返的。

定义：$\mu_i = E(T_{ii}) = \sum_{n=1}^{\infty} n f_{ii}^{(n)}$表示由状态 i 出发在返回到 i 的平均返回时间。

我们将常返态又可以分为正常返态和零常返态，设状态 i 常返$(f_{ii} = 1)$，若$\mu_i < \infty$，则称 i 为正常返；若$\mu_i = \infty$，则称 i 为零常返。
若状态 i 是正常返的，且是非周期的，则称 i 为遍历态。

$\mu_5 = \sum_{n=1}^{\infty} n f_{55}^{(n)} = 1 < \infty$ 所以状态 5 为正常返的。
我们的模型中，状态 5 是正常返的，且是非周期的，所以状态 5 为遍历态。

- **常返性的判定准则：**

$\sum_{n=0}^{\infty} p_{ii}^{(n)}$ 表示过程由 i 出发返回到 i 的平均次数

状态 i 为常返态 $\Leftrightarrow \sum_{n=0}^{\infty} p_{ii}^{(n)} = \infty$， 返回 i 的平均次数为无穷次

状态 i 为瞬时态 $\Leftrightarrow \sum_{n=0}^{\infty} p_{ii}^{(n)} = \frac{1}{1-f_{ii}} < \infty$， 返回 i 的平均次数至多为有限次

在本文的模型中，状态 0, 1, 2, 3, 4 是瞬时态，所以返回它们自身的平均次数至多为有限次。
而状态 5 为常返态，从 5 出发返回状态 5 的平均次数为无穷次。

若状态 j 为瞬时态，$\forall i \in E$, 有 $\sum_{n=0}^{\infty} p_{ij}^{(n)} < \infty$, $\lim_{n \to \infty} p_{ij}^{(n)} = 0$

若状态 j 为常返态,

- 当$i \to j$, 有$\sum_{n=0}^{\infty} p_{ij}^{(n)} = \infty$
- 当$i \nrightarrow j$, 有$\sum_{n=0}^{\infty} p_{ij}^{(n)} = 0$

状态 0，1，2，3，4 是瞬时态，所以对任意的初始状态，在转移无数次后，到达它们的概率为 0；
状态 5 是常返态，0，1，2，3，4 均可达状态 5，所以对任意的初始状态，在状态更新无数次之后，到达 5 的平均次数为无穷次。

- **零常返、正常返判定：**

设 i 常返且有周期 d，则$\lim_{n\to\infty} p_{ii}^{(nd)} = \frac{d}{\mu_i}$, μ_i为平均返回时间；当$\mu_i = \infty$时，令$\frac{d}{\mu_i} = 0$

设 i 为常返态，则：

$(1)i$为零常返态 $\Leftrightarrow \lim_{n\to\infty} p_{ii}^{(n)} = 0$

$(2)i$为遍历态 $\Leftrightarrow \lim_{n\to\infty} p_{ii}^{(n)} = \frac{1}{\mu_i} > 0$

对于状态 5，周期为 1，$\lim_{n\to\infty} p_{55}^{(n)} = \frac{1}{\mu_5} = 1 > 0$, 所以可以判断状态 5 是遍历态。

- **互通等价类：**

若$i \leftrightarrow j$, 则
（1）i 与 j 同为常返态或瞬时态；
（2）若为常返态，同为正常饭或零常返。
那么通过我们前面的分析知道，状态 1 未瞬时态，而$1 \leftrightarrow 2, 1 \leftrightarrow 3, 1 \leftrightarrow 4$，所以可以得到状态 2，3，4 均为瞬时态。

- **状态分解：**

闭集:状态空间 E 的子集 C 成为（随机）闭集，若$\forall i \in C, j \notin C$, 都有$p_{ij} = 0$

不可约：若闭集 C 中的状态都是互通的。
不可约马氏链: 若马氏链的状态空间不可约，则此马氏链称为不可约的。
不可约 Markov 链的所有状态属于同一等价类；吸收态为单一闭集。
在本文的模型中，状态 5 为吸收态，构成一个单一闭集。所以 E 含有闭子集，该过程不是不可约链。

- **判定闭集：**

C 是闭集$\Leftrightarrow \forall i \in C, j \notin C$,都有$p_{ij}^{(n)} = 0$ 即自闭集 C 的内部,不能到达 C 的外部。

i 常返，$i \to j$,则 j 必常返，且$f_{ji} = 1, i \leftrightarrow j$

说明从常返态只能到达常返态,则 E 中所有常返态组成一个闭集。
本模型中的常返态只有状态 5，它自己构成一个闭集。

马氏链具有如下性质：

- 所有常返态构成一闭集

- 不可约马氏链或者全是常返态，或者全是瞬时态。

- **分解定理：**

任一马氏链的状态空间 E，可唯一地分解成有限个或可列个不相交子集$D, C_1, C_2, \cdots$之和，使得

- 每一个C_n是常返态组成的不可约闭集；
- C_n中状态同类，任两状态互通，同为正(零)常返态，$\forall i, j \in C_n, f_{ij} = 1$;
- D由全体瞬时态组成，自C_n中的状态不能到达 D 中的状态。

那我们现在来分解本模型的马氏链：
有前面的分析知，包含 5 的基本常返闭集为$C_1 = \{5\}$, 由所以瞬时态组成的非常返集 D=$\{0, 1, 2, 3, 4\}$

下面我们来分析转移概率的极限状态与平稳分布

- **遍历定理：**

不可约马氏链，若状态都是遍历态, 则$\lim_{n\to\infty} p_{ij}^{(n)} = \frac{1}{\mu_j}$

不可约马氏链，若状态有限且非周期, 则$\lim_{n\to\infty} p_{ij}^{(n)} = \frac{1}{\mu_j}$

马氏链遍历性说明，不论从哪个状态出发，充分转移后，到达状态 j 的概率接近于一个正常数,且该常数只与 j 有关，此后记为p_j或π_j。

- **平稳分布：**

马氏链$\{X_n, n \geq 0\}$由转移概率矩阵$P = (p_{ij})_{i,j\in E}$,若存在一个概率分布$\{\pi_i, i \in E\}$满足：

$$\pi_j = \sum_{i\in E} \pi_i P_{ij}$$

则称$\{\pi_i, i \in E\}$为该链的平稳分布。也就是说, 当初始概率分布选为π后，经过 n 步转移后的概率分布仍为π。且有，有限不可约非周期马氏链必有平稳分布。
本文模型中的马氏链不是不可约的，所以没有平稳分布。

三、实证研究

我们选用一家美国公司的十一年期间（2005-2015）共 51243 条记录的历史顾客购买行为数据集作为本文的实验数据，数据给出了用户的编号(user_id)，购买金额(amount)以及购买行为发生的日期(date)。下图是数据的示例：

	user_id	amount	date
0	760	25.0	2009-11-06
1	860	50.0	2012-09-28
2	1200	100.0	2005-10-25

图 1:原始数据示例图

存在记录的客户共有 18417 个，我们先将每个用户不同年份的年度购买次数聚合起来，放入一张表中，如下图所示：

date	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
user_id											
10	1	0	0	0	0	0	0	0	0	0	0
80	1	0	1	0	1	0	1	0	1	1	1
90	1	1	1	1	1	1	1	2	1	0	0
120	0	0	0	0	0	0	0	1	0	0	0
130	1	0	1	0	0	0	0	0	0	0	0

图 2:客户购买次数

每个客户编号是唯一的，表中的数字代表此客户一年内购买产品的次数。

然后我们按照第二章第二节所述的规则来确定客户的状态，得到的客户状态的表示，表中的第(i,j)个元素表示用户 i 在第 j 年所处的客户状态，示例如下：

	0	1	2	3	4	5	6	7	8	9	10
0	1	2	3	4	5	5	5	5	5	5	5
1	1	2	1	2	1	2	1	2	1	1	1
2	1	1	1	1	1	1	1	1	1	2	3
3	0	0	0	0	0	0	0	1	2	3	4
4	1	2	1	2	3	4	5	5	5	5	5

表 1:客户状态表

设第 t 时刻处于 i 状态的客户数量为$\omega_i^{(t)}$，其中到第 t+1 时刻转移到 j 状态的客户数量为$\omega_{i,j}^{(t+1)}$,统计的时间长度为 11。则可认为状态 i 到状态 j 的转移概率$p_{ij} = \frac{1}{11}\sum_{t=0}^{10}\frac{\omega_{i,j}^{(t+1)}}{\omega_i^{(t)}}$

现在我们来计算转移概率矩阵，对于这个客户状态表，我们计算每个相邻(一步)周期之间的状态变化，从而可以计算客户在不同状态之间的转移的数量，计算后得到 11 个状态变化数量统计表。
每个周期的转移数量矩阵都加到一起整合为一个整数据集的客户转移数量矩阵，如下图所示：

	0	1	2	3	4	5
0	73515	17192	0	0	0	0
1	0	21357	18211	0	0	0
2	0	2949	0	13304	0	0
3	0	1060	0	0	10341	0
4	0	463	0	0	0	7947
5	0	720	0	0	0	17111

图 3:客户状态转移数量矩阵

通过将数据除以行和，可以计算出转移概率矩阵，如下图：

	0	1	2	3	4	5
0	0.81	0.19	0.00	0.00	0.00	0.00
1	0.00	0.54	0.46	0.00	0.00	0.00
2	0.00	0.18	0.00	0.82	0.00	0.00
3	0.00	0.09	0.00	0.00	0.91	0.00
4	0.00	0.06	0.00	0.00	0.00	0.94
5	0.00	0.00	0.00	0.00	0.00	1.00

图 4:客户状态转移概率矩阵

这就是我们在第二章模型分析中用于计算的一步转移概率矩阵 P。

我们现在可以利用实际数据计算马氏链的初始状态分布，初始的状态为 0，所以我们计算第 0 年中处于不同状态的客户数量，计算结果表明第 0 年只有两种客户状态 0 和 1，符合我们的理论，再将其除以用户的总数，得到初始时刻客户状态的频率

0 0.933485
1 0.066515
其余状态频率为 0，我们以频率近似估计概率，只保留 4 位小数后，我们得到的初始状态分布为：

$$\pi(0) = \{\pi_i(0), i \in E\} = (p_0, p_1, p_2, p_3, p_4) = (0.9335, 0.0665, 0, 0, 0)$$

又由于时齐马氏链的概率分布完全由初始分布$\pi(0)$和概率矩阵 P 决定，有$\pi(n) = \pi(0)P^n$，我们可以用这个模型来预测未来年度的顾客状态。
可以计算出初始年度 2005 年的四年后客户的状态分布：

$$\pi(0) \times P^{(4)} = \pi(0) \times P^4$$

计算后得到预测的 2009 年客户的状态分布如下：

0	0.401865
1	0.275729
2	0.128406
3	0.099567
4	0.072855
5	0.021478

真实的客户状态分布为：
0 0.514145
1 0.240159
2 0.098876
3 0.097790
4 0.032579
5 0.016452
进行误差分析，得到的均方误差为 0.052282729956662964，可以看出，本文的模型预测结果是比较准确的。用同样的方法，还可以预测未来许多年以后的客户状态。

我们还可以验证我们之前理论分析的结论，分析之前得出的如下的结论：

- **状态 0，1，2，3，4 是瞬时态，所以对任意的初始状态，在转移无数次后，到达它们的概率为 0；**
- **状态 5 是常返态，0，1，2，3，4 均可达状态 5，所以对任意的初始状态，在状态更新无数次之后，到达 5 的平均次数为无穷次。**

我们可以将 n 取一个较大的值，如 100，看看在 100 次转移过后的转移概率和客户所处的状态如何：

	0	1	2	3	4	5
0	0.0	0.0	0.0	0.0	0.0	1.0
1	0.0	0.0	0.0	0.0	0.0	1.0
2	0.0	0.0	0.0	0.0	0.0	1.0
3	0.0	0.0	0.0	0.0	0.0	1.0
4	0.0	0.0	0.0	0.0	0.0	1.0
5	0.0	0.0	0.0	0.0	0.0	1.0

图 5:100 年后的客户状态转移概率

	after 100 years
0	0.0
1	0.0
2	0.0
3	0.0
4	0.0
5	1.0

图 6:100 年后的客户状态

可以看到，用实际数据所得出的结论和我们的分析是相符的，客户无论现在处于什么状态，100 年后他一定是处于流失状态。这也与我们的常识相符，一般的产品都会更新迭代，产品推出 100 年之后一般是不会有客户去买了。

四、总结

本文通过 Markov 链建立模型，对客户关系进行了建模与研究，理论上，探究了模型的相关性质。实践上，对真实数据集进行处理，得到客户状态的转移概率矩阵的估计，并以此作为理论分析中所用的转移概率矩阵。并对未来的客户状态进行预测，得出了较为符合实际的结果。