

第一届 **太普华南杯** 数据挖掘建模竞赛

题 目.....**航空客户忠诚度与价值分析**.....

队 长.....**彭晨枫**.....

成 员.....**李立宇**.....

学校(院系).....**汕头大学计算机系**.....

指导教师.....**姜大志 孙浩军**.....

完成时间.....**2013 年 4 月 5 日**.....

综合评定成绩：_____

评委评语：

评委签名：

航空客户忠诚度与价值分析

摘 要：根据对客户保持的研究表明“顾客忠诚度提高 5%，企业利润将增加 25.125%”，建立新客户的难度是保持老客户的 5-10 倍，维系老客户的营销成本远远低于获取新客户的成本。本文基于国内某航空公司超过 60000 个会员数据，结合现有成熟的客户管理模型，在对数据样本集进行预处理后通过数理统计的方法对样本属性进行相关性分析，挑选出对构建行为忠诚度模型高影响的因子，构建客户的忠诚度价值模型。同时基于客户价值区分 RFM 分析法，建立基于均值的 RFM 模型、基于聚类的 RFM 模型和基于 PCA 处理后的 RFM 模型对客户价值水平分类，且评估三种不同的方法在客户分类上的侧重点。

关键词：客户管理；忠诚度；RFM；相关性分析；聚类；PCA；价值分类

Aviation customer loyalty and value analysis

Abstract: According to customers's research suggests that "the customer loyalty increase 5%, corporate profits will increase by 25.125%", set up the difficulty of the new customer is 5 to 10 times to keep old customers, maintain old customers marketing costs well below the cost to acquire new customers. This article is based on a domestic airline by more than 60000 members, combined with the existing mature customer management model, after preprocessing the data set by mathematical statistics method, correlation analysis was carried out on the sample properties, pick out the high influence factor for the construction of behavior loyalty model, build the loyalty of the customer value model. Based on the customer value to distinguish the RFM analysis at the same time, based on the mean of the RFM model, the RFM model based on clustering and the RFM model based on PCA after processing level of customer value classification, and evaluation of three different methods to focus on customer classification.

Key words: Customer management; Loyalty; RFM; Correlation analysis; Clustering; PCA; Value of classification

目 录

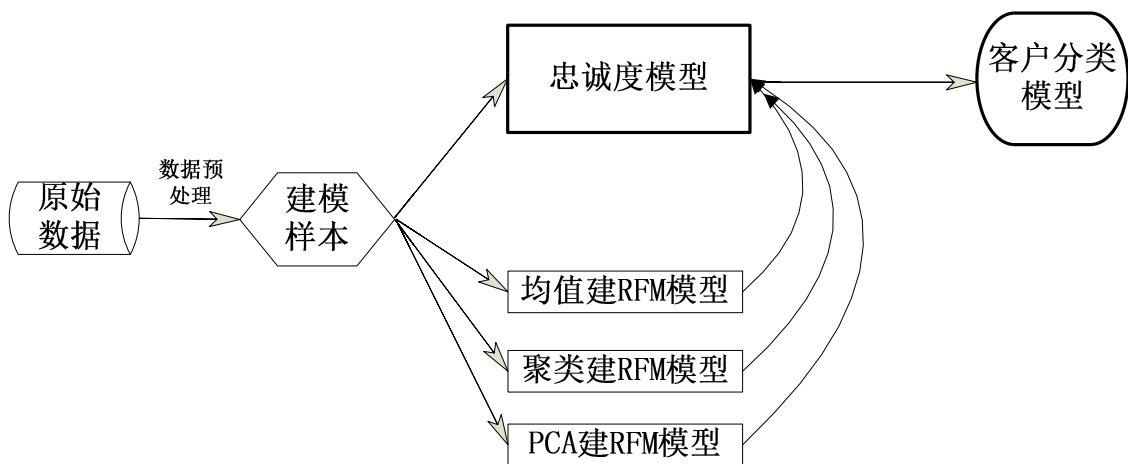
1. 研究目标.....	6
2. 分析方法与过程.....	6
2.1. 总体流程	6
2.2. 具体步骤	7
2.3. 结果分析	8
3. 结论	13
4. 参考文献.....	14

1. 挖掘目标

本次建模目标是利用太普公司提供的 62988 个航空公司的样本数据，其中包括有 63 个属性，共 8 个季度用户飞行的真实数据^[1]。对所有的数据样本进行预处理，除去缺失值的样本数据后，通过对传统客户保持管理的模型^[2]分析，结合数理统计^[3]中的一系列方法，从 63 个属性中筛选出对客户行为忠诚度^{[4][5][6]}影响高的因子，根据权衡忠诚度高影响的因子构建客户的忠诚度模型^{[7][8][9]}。同时，基于 RFM 客户价值量分析法^{[10][11]}，并结合忠诚度高影响因子构建 RFM 均值处理数据的模型、聚类^{[12][13]}处理客户数据后的 RFM 模型和 PCA^[14]降维后筛选高特征值的 RFM 模型，对客户价值水平进行分类，比较三种不同方法在客户分类上的优势与侧重点，为后期的营销策略打下前期基础。

2. 分析方法与过程

2.1. 总体流程



图（一）总体流程图

步骤一：样本数据预处理。

步骤二：Pearson 积矩系数分析，提取强相关性的因子。

步骤三：构建客户的忠诚度模型。

步骤四：建立均值 RFM 模型对客户分类。

步骤五：建立 K-means RFM 模型对客户分类。

步骤六：建立 PCA RFM 模型对客户分类。

2.2. 具体步骤

步骤 1：样本数据预处理 对构建模型不相关的属性去除，在筛选出来的属性中除去有缺失值的样本，最后挑选出 62294 个有价值的样本。

步骤 2：Pearson 积矩系数分析 利用 Pearson 方程在数值数据处理方面的优势，计算出属性之间的相关度，并结合客户管理忠诚度现有成熟的模型，对属性再次处理，筛选出对构建模型直接有效的属性样本。同时通过 Pearson 方程计算，发现客户的交易金额和积分具有强相关性，两类属性之间有超过 92% 的相关度。

步骤 3：根据已选出来的属性，分别有 3 大类：（1）客户近期交易时间；（2）客户交易频率；（3）客户交易金额。

其中，（1）客户近期交易时间具有客户距离前一次飞行间隔时间、客户平均飞行间隔、客户乘机最大时间间隔三个属性。（2）客户交易频率具有前一季飞行次数、前四季飞行次数、前八季飞行次数三个属性。（3）由步骤二的 Pearson 求相关性可知，客户交易金额与客户的积分具有强相关性。在本文中，利用客户的积分来替换客户交易金额。由此，客户积分具有前一季积分总额、前四季积分总额、前八季积分总额三个属性。

因此，客户忠诚度模型表示为： $L = T * N * I$ 。其中，L 表示忠诚度模型，T 表示客户近期交易时间，N 表示客户交易频率，I 表示客户积分，“*”表示 T、N、I 三个属性之间相同的数据样本客户之间的聚集。

步骤四：均值 RFM 模型对客户分类 传统的 RFM 模型分类是基于客户最近一次购买（R）、客户购买的频率（F）、客户购买的金额（M）三个属性来建模^{[15][16]}。基于这三个属性，本文通过求出三个属性全部样本的平均值 R^- 、 F^- 、 M^- 和每个属性中的每个样本 R_i 、 F_i 、 M_i ($i=1,2,...,62294$) 进行比较，大于均值的样本用 1 表示，小于均值的样本用 0 表示。

将客户分类定义为可保持、低价值、有潜力、要挽留四大类。

客户类型	Recency	Frequency	Monetary	客户类型
类型一	1	1	1	黄金客户
类型二	0	1	1	重要挽留客户
类型三	1	0	1	重要发展客户
类型四	0	0	1	重要保持客户
类型五	1	1	0	一般保持客户
类型五	0	1	0	一般保持客户
类型六	1	0	0	潜在发展客户
类型七	0	0	0	低价值客户

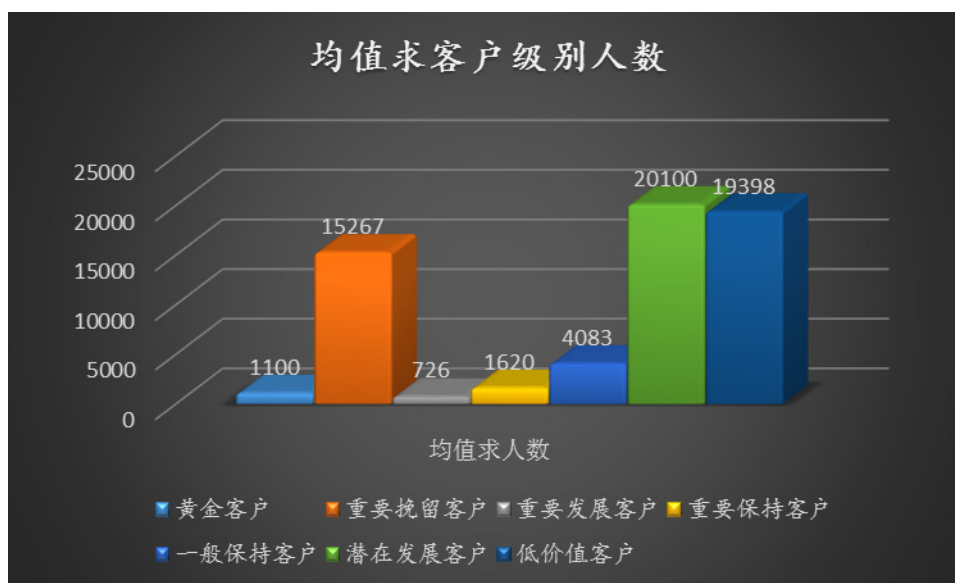
表(一) 客户价值量分类表

步骤五：K-means RFM 模型对客户分类 根据客户最近一次购买（R）、客户购买的频率（F）、客户购买的金额（M）三个属性的所有样本值通过 K-means（K=2）进行聚类，用聚类的方式求出样本之间的中心点，大于中心点的样本定义为 1，小于中心点的样本定义为 0。根据聚类后的样本进行分类（跟表（一）分类方法一样）。

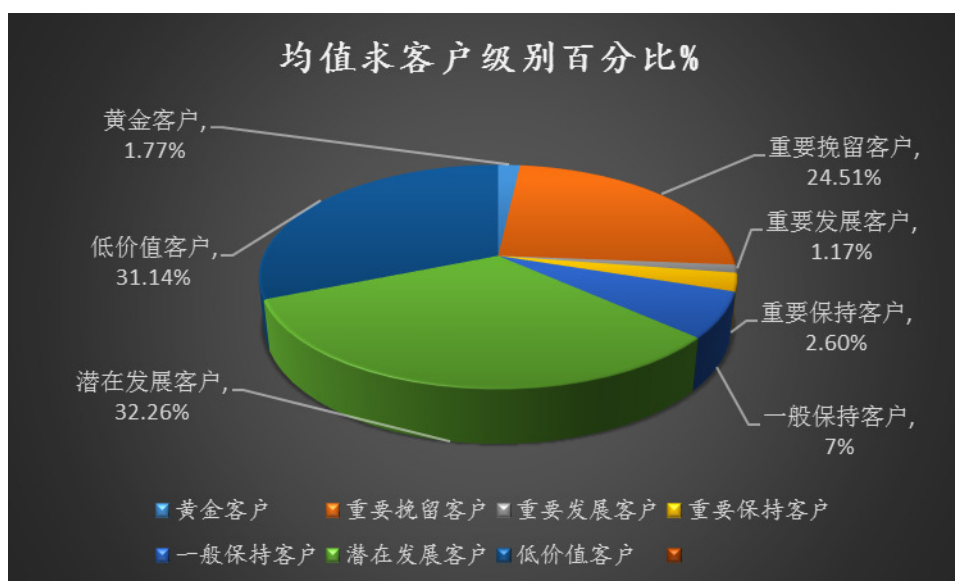
步骤六：PCA RFM 模型对客户分类 该方法突破传统 RFM 分类方式只用 R、F、M 三个属性进行建模的思想，结合 PCA（主成分分析）技术，通过步骤（3）筛选出来建立忠诚度的 3 类权衡属性：①客户近期交易时间 ②客户交易频率 ③客户积分，利用 PCA 降维的方法来构造新的 R、F、M 属性，分别求出①②③三大类属性的特征矩阵，并分别用①②③的最大特征值 y_1 与①②③的协方差矩阵建立的 62294×1 的降维矩阵来代替传统的 R、F、M 属性。并结合步骤（四）和步骤（五）的方法，对样本进行分类。

2.3. 结果分析

（1）通过对数据样本求均值的方法在客户级别上分类的分析：



图（1） 均值方法应用在求客户级别人数

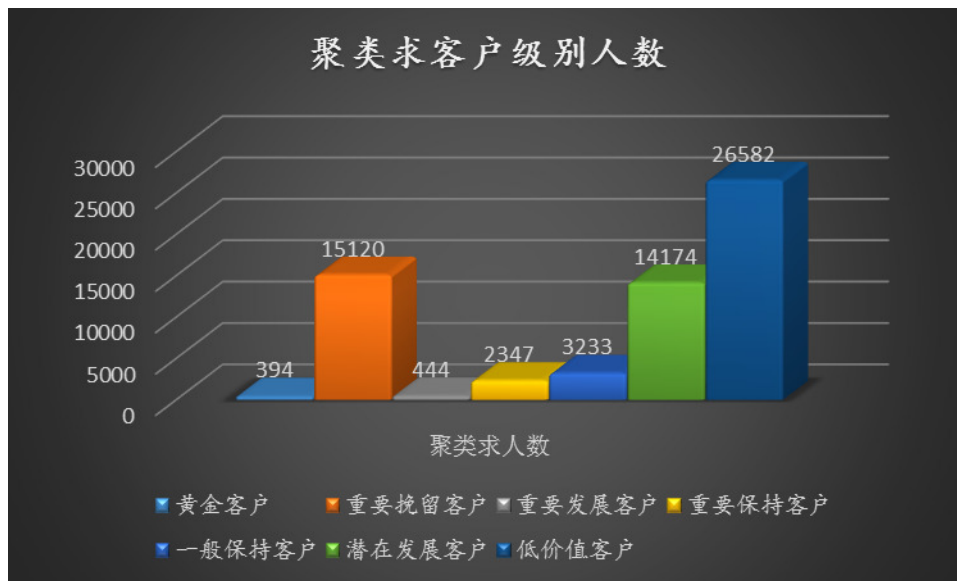


图(2) 均值方法应用在求客户级别的百分比%

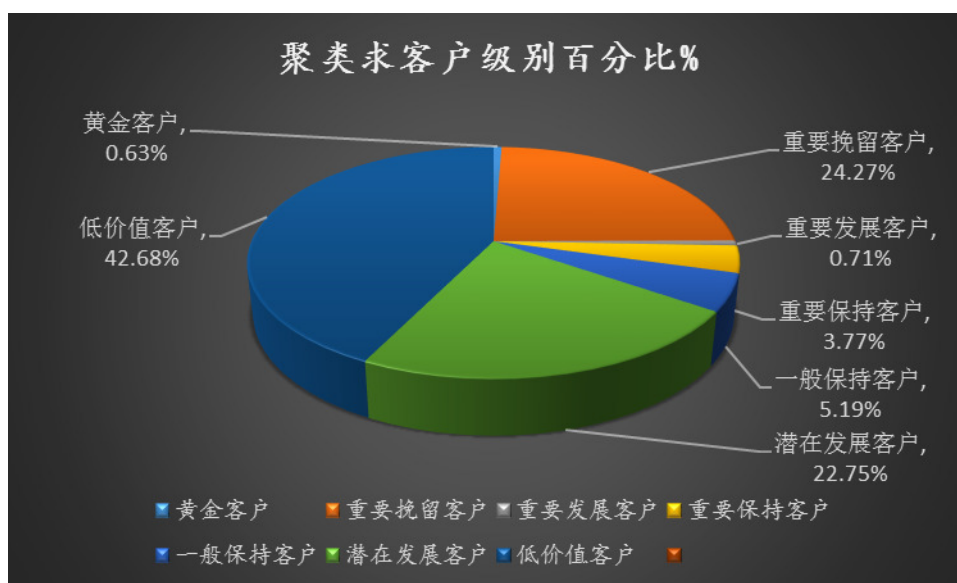
通过图(1)、图(2)可知，基于均值求客户级别的方法，航空公司 62294 个数据样本中，潜在发展客户的人数最多，占百分比的 32.26%。低价值客户居第二位，有 19398 个客户，占百分比的 31.14%。而重要挽留客户人数为 15267，占了总人数的 24.51%。最后，一般保持客户、重要保持客户、重要发展客户和黄金客户人数分别有 4083、1620、726、1100，分别占百分比 7%、2.60%、1.17% 和 1.77%。

根据 7 种不同的级别分析，潜在发展客户最多，代表了该客户具有很高的潜力，最近刚开始乘坐该航空公司的飞机，但是其消费次数和消费金额还不高，是公司的新客户，具有一定的潜力发展价值。而低价值的客户也不少，超过了总人数的三成，对公司的贡献不大。在全部的客户中，有接近 25% 的客户在历史飞行记录中，飞行次数和消费金额都很可观，但是最近没有消费，有客户流失的危险，航空公司应该对该类型的用户进行挽留，多增加其优惠政策，从客户价值角度看，该用户对公司具有重要的价值。而一般保持客户具有低消费的特征，飞行次数比较多，属于散户短途飞行的类型。重要保持客户和重要发展客户占的人数不多，但重要发展客户具有很大的潜力，最有希望可以成为该航空公司的黄金客户，公司可通过“一对一”的营销服务提高与该客户的接触，这些客户将会给公司带来更大的利润，具有重要的发展潜力价值。最后，黄金客户占的比率虽然不高，但是其的忠诚度、飞行频率和交易金额都是最高的，对公司带来了丰厚的利润，无论从客户行为还是从客户价值，该客户都属于公司高利润的源泉，公司应该重点保持和培养这类型的客户。

(2) 通过对数据样本求聚类的方法在客户级别上分类的分析：



图（3）聚类方法应用在客户级别人数上的分布



图（4）聚类方法应用在求客户级别的百分比%

通过图（3）和图（4）可知，聚类方法在对客户级别分类上，低价值客户占总人数的 42.67%，共有 26582 人。潜在发展客户人数和重要挽留客户人数相当，分别有 14174、15120 人，占总人数的 22.75%和 24.27%。一般保持客户、重要保持客户、重要发展客户和黄金客户还是占总人数的少部分，分别有 3233、2347、444、394 人，四类客户占总人数 10%左右。

(3) 通过对数据样本求 PCA 的方法在客户级别上分类的分析:

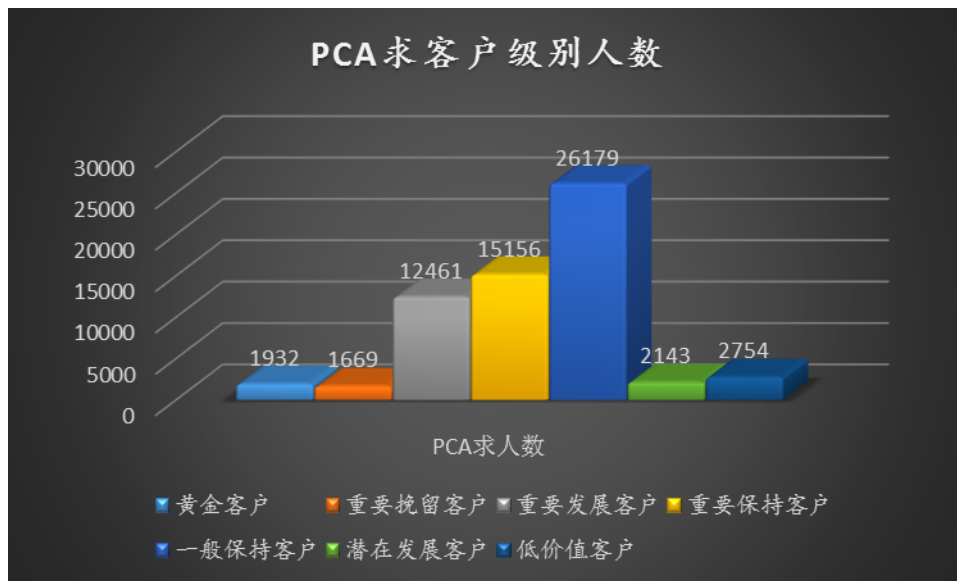


图 (5) PCA 方法应用在客户级别人数上的分布

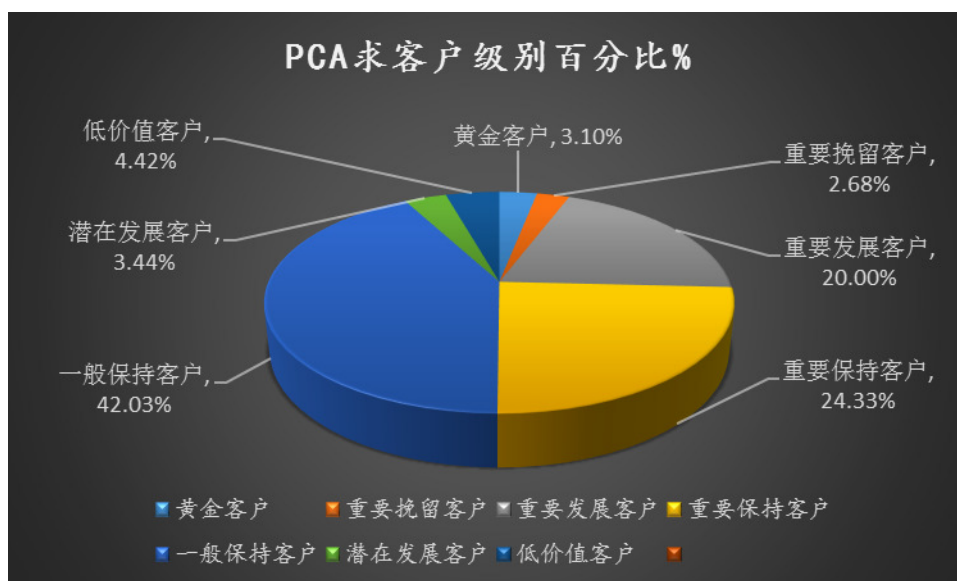


图 (6) PCA 方法应用在求客户级别的百分比%

通过 (5) 和图 (6) 可知, 在经过 PCA 处理后的数据, 在客户分类上一般保持客户占绝大部分, 共有 26179 人, 占总人数 42.03%。而重要发展客户与重要保持客户分别占 24.33% 和 20.00%, 有 15156、12461 人。潜在发展客户、低价值客户重要挽留客户和黄金客户分别有 2143、2754、1932、1669, 这四类客户人数都是占总人数比例 2% 到 5% 左右。

(4) 三种不同的方法应用在客户分类上分析:



图 (7) 三种不同方法应用在客户人数分类的分布

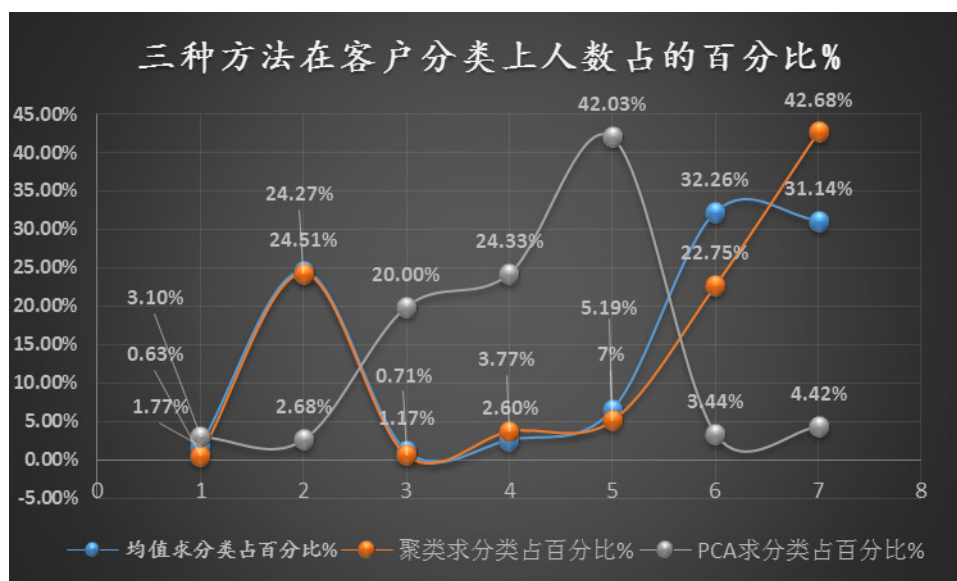


图 (8) 三种不同方法应用在客户分类人数占百分比的分布

通过 (7) 和图 (8) 可知, 三种不同颜色的线代表三种不同的方法, 蓝色为均值方法、橘黄色为聚类方法、银白色为 PCA 方法。而横坐标 1 到 7 分别代表了表 (一) 中七种不同的客户分类类型。

由人数分布图和人数所占百分比所得, 在黄金客户的分布上, PCA 方法占的人数比较多, 均值方法居中, 聚类所占人数最少。PCA 方法, 可以用来加大黄金客户的人数, 扩大营销在 VIP 级别的人数。而聚类方法, 可以压缩黄金客户的人数, 使营销方案更加具有针对性。航空公司可更加自身发展, 运用这三种不同的方法来控制黄金客户的人数。

在重要挽留客户这个类型上，聚类 and 均值的方法相当，都可以使该类型的客户分类占总人数的 25% 左右，适用于公司在发展中，更好地控制有价值客户的流失。

在重要发展客户上分类，PCA 方法具有较大的优势，该方法重视重要客户的发展，可以使重要发展的客户达总人数 20% 左右。航空公司可以用 PCA 方法，扩大发展的客户，这些重要的客户可以为公司创造很大的利润空间。

在重要保持客户的分类上，PCA 方法同样具有很大的优势，与重要发展客户分类相当，更注重重要客户的发展。适合于航空公司中期防止重要客户流失，且可以给航空公司平缓上升提供一种很好的指导方案。

在一般保持客户的分类，PCA 方法可以使该类型的人数占总人数的 42%。这种类型的客户虽然不能给公司提供较大的利润空间，但是其稳定、持续、长期与航空公司有业务交易，可以使航空公司保持盈亏在平衡点之上。

在潜在客户发展的分类上，均值和聚类的方法较 PCA 更具有优势。均值和聚类两种方法，可以用来挖掘更多有潜力，有价值的客户，可为航空公司提供充足的未来资源，为航空公司提供更大的上升发展的空间。

在低价值客户的分类上，均值和聚类两种方法都使该类型的客户占总人数 30% 到 40%。这两种方法也可以为航空公司提供参考意见，为公司的营销节约更多的成本。

总体来看，三种方法更具有各自的优点。均值和聚类两种方法在处理客户分类上，分类效果更为接近。但在操作上，聚类方法更具有价值，可根据样本平衡点控制不同类型人数的比例，排除一些特殊的用户。这两种方法有利于给刚起步的航空公司或要求有更高上升空间的航空公司而需要更多发展潜力的客户提供营销方案，对有潜力发展的客户更具有针对性，可追求更多有潜力、忠诚度不高、有价值的客户。PCA 方法，更适合已经很成熟的航空公司，这种方法更侧重于公司的平缓上升，同时可为公司保留更多的重要、有价值、忠诚度高的客户提供指导意见。

3. 结论

根据现有的客户管理方案和数据挖掘^[17]的知识对粗糙样本进行了预处理，挑选出对建模有价值的属性和样本。首先建立了一个有效的评价客户忠诚度的模型。然后基于三种不同的方法结合忠诚度模型建立客户的 RFM 分类模型，在客户分类上做了详细的分析和对比。最后，比较了三种不同方法各自的特点，为航空公司提供更有价值的营销方案。

4. 参考文献

- [1]ftp://221.234.10.59/questions.zip
- [2]朱爱群. 客户关系管理与数据挖掘[M]. 北京: 中国财经经济出版社. 2001: 184-212.
- [3] K Pearson. Mathematical contributions to the theory of evolution. VII. On the correlation of characters not quantitatively measurable. Philosophical Transactions of the Royal Society of London. 1900.
- [4] Buchanan, R. Gilles, C. Value managed relationship: The key to customer retention and profitability. European Management Journal. 1990: 8(4).
- [5] Dawkins, P. and Reichheld, F. Customer retention as a competitive weapon. Directors and Boards. 1990: 14(4).
- [6]Carrol, P. Reichheld, F. The fallacy of customer retention. Journal of Retail Banking. 1992: 13(4).
- [7]Bult, J. R.. Wansbeek, T. J. Optimal selection for direct mail. Marketing Science. 1995: 14(4): 378-394.
- [8]周贺来. 客户关系管理实务. 北京: 北京大学出版社[M]. 2011: 100-128.
- [9]曹洁慧. 基于数据挖掘的我国证券业客户忠诚度研究[D]. 2007.
- [10] YL Chen. MH Kuo. SY Wu. K Tang. Discovering recency, frequency, and monetary (RFM) sequential patterns from customers' purchasing data. Electronic commerce research and applications. 2009.
- [11]邵兵家. 客户关系管理(第二版). 北京: 清华大学出版社. 2004: 52-61.
- [12] J. A. Hartigan. Clustering Algorithms. Wiley: 1975.
- [13] J. A. Hartigan. M. A. Wong. A K-Means Clustering Algorithm. Applied Statistics. 1979: 28(1): 100-108.
- [14] Pearson, K.. On Lines. Planes of Closest Fit to Systems of Points in Space. Philosophical Magazine. 1901: 2(6): 559 - 572.
- [15]韩锦. 邹云峰. 高华. 基于 RFM 模型的聚类分析算法在期货公司客户分类中的应用. 长江期货. 2012.
- [16] Arthur Hughes. 电子商务网站 RFM 分析客户关系. 美国数据库营销研究所. 2010.
- [17] Jiawei Han. Micheline Kamber. Data Mining Concepts and Techniques Third Edition[M]. America: Morgan Kaufmann Publishers, 2011.