

基于改进LRFMC模型的航空公司客户分类

李莹^{1*}, 颜轲越²

¹北京理工大学珠海学院, 中美国际学院, 广东 珠海

²澳门大学科技学院, 澳门

收稿日期: 2022年4月18日; 录用日期: 2022年5月17日; 发布日期: 2022年5月24日

摘要

随着人们生活水平的提高, 越来越多的人会选择飞机交通工具出行。对于同一个航线, 人们可以从多个航空公司的航班之间进行选择, 这一行为使得各个航空公司的竞争不断加剧。如何保留住老客户的同时, 还增加新客户数量成为了各个航空公司需要解决的问题, 客户关系管理是其基础任务。客户关系管理首先需要将客户根据不同特性划分成多个类别。本研究将基于国内某航空公司客户的数据对LRFMC模型进行调整和改进。为了达到降低LRFMC模型中指标的共线性和完善的目的, 将原模型中的飞行总公里数这一指标替换成平均每次飞行公里数并增加每公里机票票价构建新模型。使用层次分析法计算新模型中各指标的权重, 最后通过K-Means算法对客户数据进行分类。根据分类的结果, 航空公司通过对不同价值的客户采取不同的方式和策略来提高其公司的收益。

关键词

客户分类, LRFMC模型, 层次分析法, K-Means算法

Airline Customer Classification Based on Improved LRFMC Model

Ying Li^{1*}, Keyue Yan²

¹College of Global Talents, Beijing Institute of Technology (Zhuhai), Zhuhai Guangdong

²Faculty of Science and Technology, University of Macau, Macau

Received: Apr. 18th, 2022; accepted: May 17th, 2022; published: May 24th, 2022

Abstract

With the improvement of the level of life, more and more people will travel by airplane. For the same route, people can choose it from multiple airlines' flights, which makes the competition

*通讯作者。

文章引用: 李莹, 颜轲越. 基于改进 LRFMC 模型的航空公司客户分类[J]. 计算机科学与应用, 2022, 12(5): 1341-1349.
DOI: 10.12677/csa.2022.125133

among airlines. How to keep customers and increase the number of customers has been a problem that airlines need to solve, and customer relation management (CRM) is the basic task. CRM needs to classify customers into multiple categories based on their characteristics. This study will adjust and improve LRFMC model based on the data of airline customers. In order to reduce the covariance in LRFMC model, the new model is constructed by replacing the index of total kilometers with average kilometers per flight and adding average fare per kilometer. The weights of each indicator in the new model are calculated by using Analytic Hierarchy Process (AHP), and finally the customers' data are classified by K-Means algorithm. Based on the results of classification, airlines can adopt different approaches and strategies for customers with different values to improve revenue.

Keywords

Customer Classification, LRFMC Model, Analytic Hierarchy Process, K-Means Algorithm

Copyright © 2022 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

由于当今航空行业的发展和受疫情的影响,人们对于远距离的出行会优先考虑飞机这一交通工具。在缩短出行所需时间的同时,还可以减少旅途中感染病毒的风险。对于任何一个公司企业,客户都是所有环节中基础且重要的组成部分。航空公司为了更好的发展,需要探究其客户的价值并针对不同等级价值的客户采取不同的营销方式和措施[1]。

文献[2]改进原有的传统 RFM 模型并提出 LRFMC 模型用来解决航空公司客户的价值分类问题,考虑到消费金额不能完全表示客户的消费能力。包志强和赵媛媛等[3]对传统的 RFM 模型进行改进,使用 RFMS 模型对百度外卖的客户价值进行分析,增加对新老客户的考虑,同时发现各指标之间存在线性关系而采用主成分分析法来确定指标权重。冀慧杰和倪枫等[4]运用 RFMT 模型对电子商务客户群体进行划分,为了挖掘各指标间的关联性,使用灰色关联度得到各指标的权重,分类结果因此变得更准确。陈燕萍和王文杰[5]基于 RFMD 模型对健身行业的客户的价值分类,利用肘部法则确定客户的聚类个数。黄先林[6]提出了 RFMF 模型并根据该模型划分了公募基金用户的类别,为了更好地确定聚类个数而使用肘部法则和 CH 指标两种方式。周洋颀[7]使用 DRFKM 模型对航空公司的客户进行价值分析并利用逻辑回归法预测了客户的流失。

结合上述文献的不足与改进,本研究基于 LRFMC 模型,提出 LRFPMD 模型并对航空公司的客户进行分类,使用国内某航空公司客户真实的历史数据作为实验数据集。首先对数据集进行预处理,然后提取模型中所需的相关变量并对相关性较高的变量进行调整。由于各指标之间的共线性问题已被解决,同时需要考虑到不同指标对客户价值的重要性不同,所以使用层次分析法[8]来确定各指标权重。最后使用 K-Means 算法对模型进行分类,利用肘部法则和轮廓系数法[9],能够更准确地确定客户聚类个数。

2. 研究方法

2.1. 数据描述与预处理

本研究使用的是在 Heywhale 平台[10]下载的国内某航空公司客户的历史数据作为数据集。该数据集

包含客户的基本信息、观测窗口开始到结束期间内的乘机信息和积分信息, 共计 62,988 个样本, 44 个变量。其中观测窗口时间为 2012 年 4 月 1 日至 2014 年 3 月 31 日。

在 44 个变量中, “FFP_DATE”、“FIRST_FLIGHT_DATE”、“LOAD_TIME”和“LAST_FLIGHT_DATE”表示的是该航空公司客户的入会日期、第一次飞行日期、观测窗口结束的日期和观测窗口结束前的最后一次飞行日期。根据这些变量可以判断该客户是否为老会员, 近期是否乘坐该航空公司的航班。

“FLIGHT_COUNT”、“SUM_YR_1”、“SUM_YR_2”、“SEG_KM_SUM”、“WEIGHTED_SEG_KM”和“avg_discount”表示的是客户在观测窗口期间的总飞行次数、第一年的总机票票价、第二年的总机票票价、总飞行公里数、总加权飞行公里数和平均折扣率, 代表着客户的消费能力。本研究是对客户在观测时间内的整体价值进行分析讨论, 因此将“SUM_YR_1”和“SUM_YR_2”求和得到新变量“SUM_YR”, 即为观测窗口期间的总机票票价。

通过观测数据集, 可以得到各变量的取值范围和空值数量。其中“WORK_CITY”和“WORK_PROVINCE”的空值较多, “WORK_COUNTRY”有少量空值。这三个变量代表客户的工作地点, 对客户价值分类的影响并不大, 所以可以直接删除这些变量。同时, 本研究并不考虑客户的积分信息及其使用情况, 所以删除所有积分信息及兑换情况的变量, 可以减小处理数据时的储存空间和加快运算速度。“SUM_YR_1”和“SUM_YR_2”存在少量的空值, 需要删除其所在的样本。对于取值范围, 观测到“SEG_KM_SUM”最小值不为 0, 但新变量“SUM_YR”最小值为 0, 表示的是该客户在观测时间内没有花费任何金额就乘坐了航班, 此情况可能为使用里程数或积分兑换航班或者异常数据, 删除其所在的样本。除此之外, “AGE”表示客户的年龄, 其最大值为 110, 为了更直观判断该值是否为异常值, 该变量是否还存在其他异常值, 需要将其分布使用箱型图绘制出来。其分布如下图 1 所示:

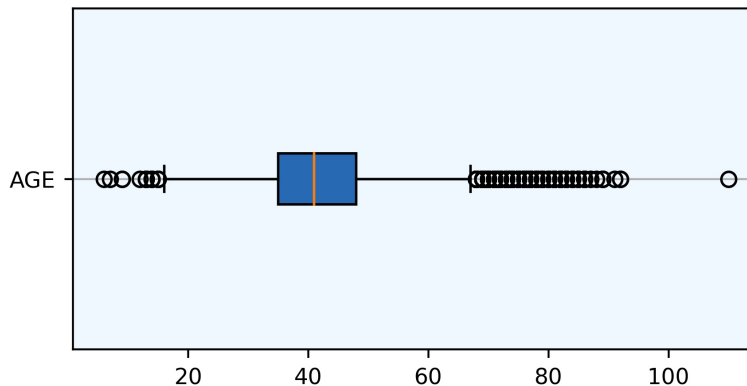


Figure 1. Distribution of the customers' age
图 1. 客户年龄分布图

由上图可知, 客户的年龄集中在 30~50 之间, 表示该航空公司的主要客户为中年群体。少数数据分布在 20 以下和 70~100 之间并且相邻两值的差异不大。由于只有一个数据 110, 也就是该变量的最大值, 在 100 以上且与其他数据有较大差异, 所以可以将其所在样本删除。

预处理后的数据样本为 61,631 个, 总共删除的样本不到原始数据的 2.2%, 直接删除对分类结果的影响不大。

2.2. 数据分析

根据“FFP_DATE”该变量中的年份, 绘制出年份与入会人数的关系。其分布如下图 2 所示:

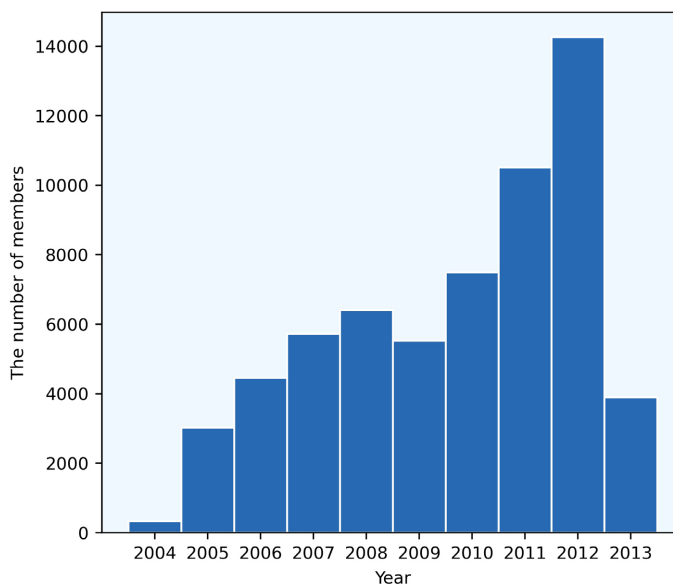


Figure 2. Distribution of the number of members each year

图 2. 各年份客户入会人数分布图

由上图可知,从 2004~2013 年每年均有客户成为会员,说明该航空公司的总会员人数每年都在增加。除 2013 年外,每年的入会人数大致呈增长趋势,在 2012 年达到最高值,说明该航空公司的入会人数的增长速度也在增加。在观测窗口开始到结束期间,即 2012~2014 年,有不少客户成为会员并且在观测窗口结束之前乘坐航班。在总机票票价相同时,在观测期间内才入会的客户要比早期入会的成员更具有价值,该变量对客户价值的分类具有一定意义。

对数据集中部分连续性变量进行相关性分析,绘制其相关系数矩阵的热力图,如下图 3 所示:

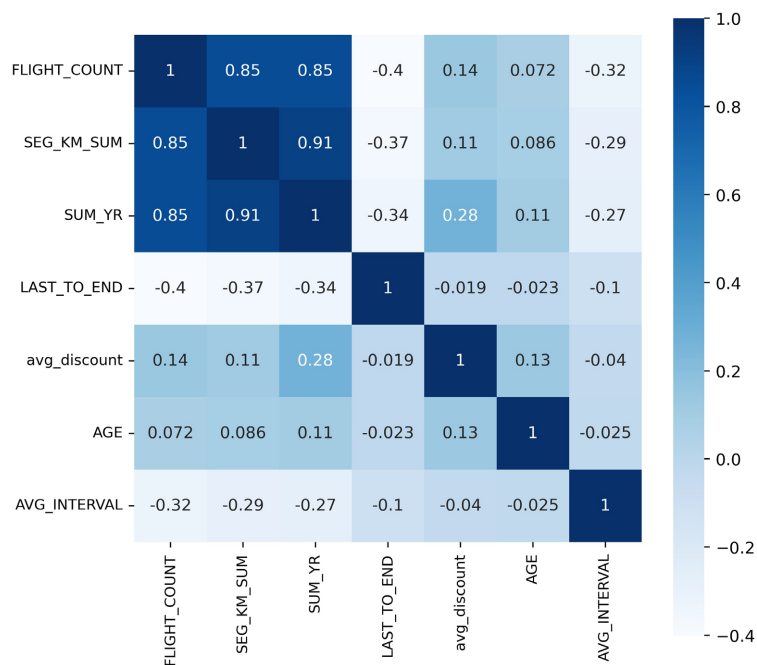


Figure 3. Correlation matrix of variables

图 3. 变量的相关系数矩阵

“FLIGHT_COUNT”、“SEG_KM_SUM”和“SUM_YR”三者之间具有很强烈的正相关性, 两两之间的相关系数分别为 0.85、0.91 和 0.85。但由于这三个变量对研究客户的价值都具有一定意义, 因此不能将其中任意一个变量直接删除。需要将变量进行调整, 来降低这三者之间的相关性。除此之外, 各变量之间的相关系数都比较低。

3. 改进 LRFMC 模型

RFM 模型经常被用于客户的分类及其价值的分析。其中 R 表示客户最近一次的消费时间(Recency), F 表示客户的消费频率(Frequency)以及 M 表示客户的消费金额(Monetary) [11]。客户的消费时间距离观测结束越近说明该客户近期越活跃, 当公司对其提供服务时得到的反馈可能会越快。客户的消费次数和金额越多说明该客户可以给公司带来较大的收益, 该客户对公司的价值更高。通过这三个指标可以将客户进行分类, 并对不同类别的客户提供相应的服务和措施。基于 RFM 模型, 不同行业对模型的指标进行重新选择或者改进。对于航空行业, 有文献提出将原指标消费金额 M 替换为总飞行里程数 M 和折扣系数平均值 C, 添加新指标客户关系长度 L, 此模型命名为 LRFMC [2]。

本研究基于 LRFMC 模型进行改进, 其中客户的总飞行公里数和折扣系数并不能完全替代总机票票价。对于短距离飞行但选择高等舱位的客户, 其价值对于航空公司要更高。在数据集中“SUM_YR”和“avg_discount”的相关性并不高, 说明平均折扣率高并不能表示该类客户, 所以本研究添加总机票票价这一指标, 记为 P。其余指标均保留, 新模型记为 LRFPM D。因为数据集中的“FLIGHT_COUNT”、“SEG_KM_SUM”和“SUM_YR”三个变量之间的相关性过高, 需要对其进行处理来降低共线性。在 LRFPM D 模型中, 构造的指标如下:

L: 入会时长, 即航空公司客户成为会员的时间到观测窗口的结束时间的间隔天数。

$$L = \text{LOAD_TIME} - \text{FFP_DATE} \quad (1)$$

R: 最近乘机时间间隔, 即距离观测窗口结束最近的一次乘机时间到结束时间的间隔天数。

$$R = \text{LAST_TO_END} \quad (2)$$

F: 观测时间内的总飞行次数。

$$F = \text{FLIGHT_COUNT} \quad (3)$$

P: 观测时间内平均每公里的机票票价, 即观测窗口时间内的总机票票价除以总加权飞行公里数。

$$P = \text{SUM_YR} / \text{WEIGHTED_SEG_KM} \quad (4)$$

M: 观测时间内平均每次飞行的公里数, 即观测窗口时间内的总飞行公里数除以总飞行次数。

$$M = \text{SEG_KM_SUM} / \text{FLIGHT_COUNT} \quad (5)$$

D: 观测时间内客户乘坐航班的平均折扣系数。

$$D = \text{avg_discount} \quad (6)$$

从数据集中提取 LRFPM D 模型所需的变量并进行运算整理, 得到各指标的数据范围如表 1 所示。

通过观察表 1 的数据, 发现各指标的取值范围差异过大, 为了减小各指标数值对分类结果的影响, 需要对其进行标准化处理来统一量纲。为了保留原本的数据分布, 本研究采用的是 z-score 标准化的方式。同时在 LRFPM D 模型中, 指标 R 越小表示客户最近乘机时间间隔越小, 近期活跃度越高, 其价值越高, 所以将指标 R 转换成 1/R 再进行客户分类。

Table 1. Data range of LRFPMO indicators
表 1. LRFPMO 模型指标数据范围

指标	平均值	标准差	最小值	最大值
L	1485.0970	847.3269	365	3437
R	172.3647	181.4881	1	731
F	11.9936	14.1333	2	213
P	0.8902	0.2733	0.0221	8.4127
M	1583.3879	1128.1624	184	13,320.5
D	0.7220	0.1847	0.1360	1.5

4. 实验

4.1. 指标权重

由于 LRFPMO 模型中各指标之间的相关性极低, 所以对于各指标权重的确定采用的是层次分析法, 该方法可以明确各指标之间的重要性。层次分析法(Analytic Hierarchy Process, AHP)是由美国萨蒂教授提出的一种将目标问题通过定性量化的方式进行分析决策权重的方法[8]。AHP 需要将目标问题层次化, 分为目标层、准则层和方案层, 再将各指标根据重要性或者主次程度进行划分并排序。通过两两指标之间比较其重要程度的方式, 对指标进行量化。将两个指标互换位置后, 其量化值则为原量化值的倒数。将所有指标通过两两比较的方式得到的值填入矩阵中, 该矩阵为判断矩阵。该判断矩阵是否成立, 需要进行一致性检验。计算一致性指标(Concordance Index, CI)与随机一致性指标(Random Index, RI)的比值得到一致性比率(Concordance Rate, CR)。当 CR 小于 0.1 时, 判断矩阵通过一致性检验; 否则需要对各指标的重要性进行重新排序来调整判断矩阵, 直到满足条件为止[12]。其公式如下所示:

$$CI = (\lambda - n) / (n - 1) \tag{7}$$

$$CR = CI / RI \tag{8}$$

其中 λ 为判断矩阵的最大特征值, n 为判断矩阵的阶数。

在 LRFPMO 模型中, 将各指标的重要性进行排序并对其量化, 得到下表 2 所示的判断矩阵。

Table 2. Judgment matrix of LRFPMO
表 2. LRFPMO 模型的判断矩阵

判断矩阵	L	1/R	F	P	M	D
L	1	1	1/3	1/4	1/5	1/2
1/R	1	1	1/3	1/4	1/5	1/2
F	3	3	1	1/2	1/3	2
P	4	4	2	1	1/2	3
M	5	5	3	2	1	4
D	2	2	1/2	1/3	1/4	1

通过计算, 该判断矩阵的最大特征值为 6.0808。通过计算得到 $CI = 0.0162$, $CR = 0.0130 < 0.1$, 其中 $RI = 1.24$, 说明该判断矩阵通过一次性检验。最大特征值对应的特征向量为(0.1176, 0.1176, 0.3145, 0.4988, 0.7661, 0.1939), 经过归一化处理后, 得到各指标的权重为 L (0.0585), 1/R (0.0585), F (0.1566), P (0.2484), M (0.3814)和 D (0.0966)。

4.2. 聚类算法

本研究针对客户价值分类问题使用的是常见的 K-Means 算法, 该算法较于其他聚类算法简单且易实现。K-Means 算法是将相似的样本归为一类。首先要确定聚类个数 K 值和其聚类中心, 将每一个样本分配到最相近的聚类中心, 得到新的聚类中心。算法将一直重复此过程, 直到均方差开始收敛[13]。

确定聚类个数的方法采用的是肘部法则和轮廓系数法相结合的方式[9]。肘部法则需要计算各样本到聚类中心的误差平方和(Sum of squared error, SSE)。随着 K 值的增大, SSE 越小, SSE 下降速度逐渐减慢。当 SSE 曲线出现“拐点”时, 为最佳 K 值。轮廓系数法是减小同类之间的差异的同时, 增大不同类别的差异。轮廓系数的取值范围为-1 到 1, 其值记为 s。s 越接近于 1, 说明分类效果越好; 反之, s 越接近 -1, 说明聚类效果越差。

由图 4 可知, 当 K 值从 1 到 2 时, SSE 快速下降; 由 2 到 7 时, SSE 下降速度逐渐减慢; 大于 7 后, 趋于平缓。但“拐点”并不明显, 无法确定最佳 K 值。由图 5 可知, K = 2 时, 轮廓系数达到最大。但因为本研究的样本数据较大, 当样本数据分为两类时, 结合图 4, SSE 较大, 分类效果较差。除去 K = 2, 轮廓系数最大时, K 值为 6, 同时也满足肘部法则的最佳 K 值范围。

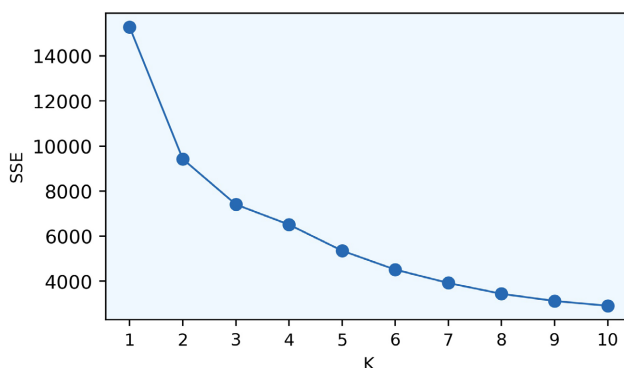


Figure 4. Determine the value of K by Elbow Rule

图 4. 肘部法则确定 K 值

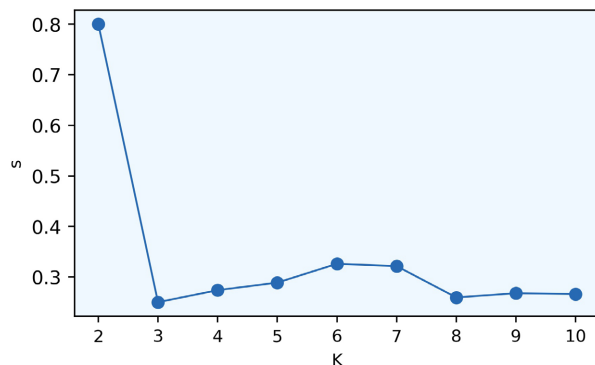


Figure 5. Determine the value of K by Silhouette Coefficient

图 5. 轮廓系数法确定 K 值

4.3. 实验结果分析

使用 Python 编程语言调用 sklearn.cluster 工具箱中 KMeans 函数[14], 将 LRFPMD 模型中的各指标作为聚类变量进行分类。为了便于观测各类客户的特点, 将聚类后的类别放入处理前的 LRFPMD 模型中, 得到的聚类结果如下表 3 所示:

Table 3. Clustering results of LRFPMD
表 3. LRFPMD 模型的聚类结果

类别	聚类个数	聚类中心					
		L	R	F	P	M	D
1	5556	1913.8963	27.3228	46.6758	0.9142	1290.9265	0.7944
2	981	1349.3129	219.3384	5.9011	0.5042	8639.4949	0.7475
3	7463	1444.6115	186.6857	9.5965	0.5145	1347.6370	0.7195
4	754	1212.2877	251.4721	9.6392	2.3601	823.5337	0.8032
5	33,213	1456.2196	178.7312	8.7456	0.9768	185.5210	0.7264
6	13,664	1427.8473	200.3061	7.6624	0.8216	2333.5023	0.6766

由上表可知, 类别 1 的 L 和 F 是所有类别里最高的, P 和 D 也较高, R 最低, 其中 R 和 F 和其他类别的存在明显差异。说明该类客户经常需要乘坐飞机出行, 且入会时间久, 近期乘坐该航空公司的航班, 其航班目的地距离中等。该类别的客户是该航空公司的主要客户, 其忠诚度很高, 也是航空公司收益的主要来源。但该类客户的人数仅是总实验样本数的 9%, 需要航空公司对其进行单独管理, 同时采取相应的措施来吸引新的客户, 带来更多的利益。

类别 2 的 M 是所有类别里最高的, R 较高, 但其它指标的值较低, 尤其 F 和 P 是最低的。说明该类客户飞行次数少, 平均飞行距离远, 但每公里的机票票价过低且已经很久没有乘坐该航空的航班出行。该类客户可能是购买廉价的低等舱位机票, 可能已经改换其他航班或者其他交通方式出行, 是该航空公司流失客户。

类别 3 的 P 和 D 较低。说明该类客户可能经常购买低等舱位的廉价机票或者参与机票的打折活动。虽然近期没有出行, 但该类客户飞行次数并不低。该类客户属于该航空公司的低价值客户。

类别 4 的 R、P 和 D 是所有类别里最高的, 同时 L 是最低的。说明该类别的客户入会时间较短, 出行的目的地距离较近, 但乘坐高等舱位且折扣较小且已经很长时间没有乘坐该航空公司的航班。该类客户是该航空公司主要挽留客户, 其潜在的价值要高于主要客户。需要航空公司对其进行联系, 了解其需求, 增加该类客户的忠诚度, 提高其活跃度和飞行次数。

类别 5 的 M 是所有类别里最低的, L 和 P 较高, R 较低。说明该类客户入会时间久, 近期仍在使用该航空公司的航班出行, 平均飞行距离过短, 参与过打折活动, 购买高等舱位的机票。该类客户人数最多, 但属于一般客户。虽然每公里机票票价较高, 但出行距离太近, 飞行次数也不够多, 能够带给航空公司的收益有限。

类别 6 的 D 是所有类别最低的, M 较高, F 较低。说明该类客户的平均飞行距离较远。虽然折扣系数很低, 但每公里的机票票价不低, 可能购买的是打折的高等舱位。该类别的客户属于该航空公司的发展客户, 且人数较多。需要该航空公司增加其活跃度, 提高客户的飞行次数, 增加该类客户的价值。

5. 总结

本研究针对航空公司的客户价值分类的问题,使用的是 K-Means 算法。首先对 LRFMC 模型进行改进,增加一个新指标平均每公里的机票票价 P 。通过分析数据集,发现模型中部分指标的相关性较高,便将相关性高的指标进行调整来降低彼此之间的相关系数。本研究对于模型中各指标权重的确定采用的是层次分析法,再使用 K-Means 算法对数据集进行分类。其中对于 K 值的确定使用的是肘部法则和轮廓系数法相结合的方式。这样既可以降低聚类时的误差,又可以使不同类别的差异性增大。最后根据分类结果,对各类别的客户特征进行分析。

参考文献

- [1] 孙文仪. 浅谈企业如何实施客户关系管理[J]. 上海商业, 2021(12): 104-105.
- [2] 张良均. Python 数据分析与挖掘实战[M]. 北京: 机械工业出版社, 2016: 164-179.
- [3] 包志强, 赵媛媛, 赵研, 等. 基于改进 RFM 模型的百度外卖客户价值分析[J]. 西安邮电大学学报, 2019, 24(1): 105-110.
- [4] 冀慧杰, 倪枫, 刘姜, 赵隼. 基于灰色关联度和 K-Means++ 的电子商务客户价值分类[J]. 计算机系统应用, 2020, 29(9): 249-254.
- [5] 陈燕萍, 王文杰. 基于 RFMD 模型和聚类分析的健身休闲行业客户分类[J]. 电子商务评论, 2021, 10(2): 25-36.
- [6] 黄先林. 基于 RFMF 模型和 K-Means 方法的公募基金用户精细化营销[J]. 金融, 2019, 9(6): 634-640.
- [7] 周杨嶝. 航空公司客户价值分析及流失预测[J]. 商讯, 2021(18): 190-192.
- [8] 许树柏. 实用决策方法: 层次分析法原理[M]. 天津: 天津大学出版社, 1988: 1-5.
- [9] 和志强, 王梦雪, 马宁, 等. 短文本聚类方法研究综述[J]. 河北省科学院学报, 2021, 38(5): 34-40.
- [10] Heywhale (2020). <https://www.heywhale.com/home/dataset>
- [11] 王文贤, 金阳, 陈道斌. 基于 RFM 模型的个人客户忠诚度研究[J]. 金融论坛, 2012(3): 75-80.
- [12] 许树柏. 实用决策方法: 层次分析法原理[M]. 天津: 天津大学出版社, 1988: 43-68.
- [13] 郑志明, 缪绍日, 荆丽丽. 金融数据挖掘与分析[M]. 北京: 机械工业出版社, 2016: 59-62.
- [14] KMeans (2022). <https://scikit-learn.org/stable/modules/clustering.html#k-means>