商务数据分析期末报告：航空会员特征研究

清华大学经济管理学院　王天宇

2020 年 6 月 21 日

摘要与前言

信息时代下航空市场竞争日益加剧, 国内航空公司更需要利用这些大量的客户数据发现其中隐含的规律和模式，进而提升航空公司的竞争力。航空公司可以在客户关系管理中使用数据挖掘技术，对客户价值进行分类建模, 这样能够使为其发展提供决策和经营的参考基础, 从而扩大航空公司的用户基础并提高航空公司利润空间。在航空公司会员原始数据的基础上, 首先我们按照业务逻辑对其从异常值, 缺失值等角度进行预处理。然后对于客户的基本信息和飞行记录等信息，我们从数据分布, 不同数据之间的相关性和分组关系进行数据探索, 并给出相应的可视化结果。

在此基础上，我们建立了基于 LRFMC 的客户价值识别模型和客户流失模型。我们首先使用 KMeans 聚类方法将客户价值分为不同类别。然后使用常用的机器学习方法对于客户价值进行了预测，而对于老客户的客户流失模型, 我们通过老客户的乘机次数进行分类判断并给出预测，均达到了较高的预测精度。因此，航空公司可以根据飞行次数, 频率, 积分等特征制定相应的营销策略, 为不同类别的用户给出相应的差异化服务，更好地挖掘他们的潜在价值。

关键词：数据挖掘　客户价值　客户流失

目录

1 问题介绍 **3**

2 数据基本介绍与预处理 **3**

3 数据统计分析 **5**

3.1 用户特征画像 5

3.2 用户飞行行为画像 10

4 数据挖掘分析 **14**

4.1 客户端与企业端心智模型的交互 14

4.2 客户价值识别模型 14

4.3 客户流失模型 19

5 商务分析与应用前景 **22**

6 总结与建议 **23**

7 附录 **25**

1 问题介绍

电子商务正在猛烈冲击着传统企业与客户的关系，随着大数据分析技术的兴起，面对各行各业产生的海量数据，企业拥有了精确分析客户需求并从中在市场上更快反应的能力，以此对客户更好地实现差异化定价并扩大其客户规模。针对实际问题，企业遇到的核心问题便是客户分类识别，进而针对不同类别的客户制定针对性的个性化服务。这便是**客户关系管理 (Customer Relationship Management**, [1]) 的核心。这样在企业为不同顾客不同的营销管理策略之后，才能使得客户的价值更好地流向企业，同时企业的服务能够更有针对性地流向客户，实现双方的互利共赢 ([2])。

21 世纪以来，中国民用航空业得到了迅速发展，各家航空公司在市场上也竞争较为激烈。尽管航空公司很早开始采用收益管理的策略进行动态定价，但是这种粗放型经济没有结合用户个人的身份特征，从而造成较大的资源损失。航空公司的传统信息系统包含大量客户特征和飞行记录信息的今。如何使用信息技术使得这些数据更好地为客户服务，成为了新时代航空公司竞争的主要方向。

因此，本课题致力于从航空公司原有的会员记录出发，探究其会员的信息特征、购票行为和积分信息等特征与客户行为的关系，进而从基础统计量分布、数据可视化、数据分类建模学习等多种数据挖掘工具，识别在这一给定数据集中用户的价值模型，进而对航空公司进一步留住、发展会员等提供决策上的参考。

本课题针对航空公司客户价值分析的总体流程如下：

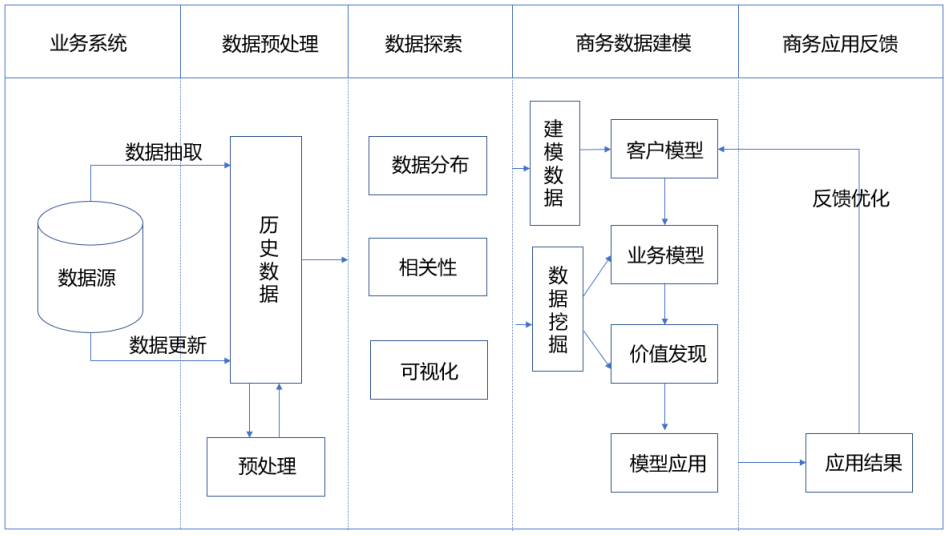

图 1: 客户价值建模

2 数据基本介绍与预处理

在基本的数据探查之后，原始数据集共有 **62988 条**，其中包括 63 个特征。**本数据为航空公司抽取 2015 年 4 月 1 日到 2017 年 3 月 31 日 (共有八个季度) 的会员的详细数据作为观**

测窗口的信息。该数据集的主键为会员卡号 (MEMBER_NO), 按照入会时间进行排序。经过初步探查后其中不包括重复项, 而根据原始数据的字段描述，可将其分为如下三类：

表 1: 航空公司客户特征属性表

分类序号	字段条目	信息分类	主要字段
1	1-10	会员基本信息	入会时间, 年龄, 会员卡级别
2	12-20, 32-43, 47-50	会员乘机信息	飞行次数, 第一年总票价, 平均乘机时间间隔
3	21-31, 44-46, 51-63	会员积分信息	观察窗口总积分信息, 总累计积分，非乘机积分总和

在此基础上，我们对数据进行预处理的探查和分析。我们发现在 63 个字段中，共只有 8 项数据存在空值。数据集字段的数据类型主要集中在整型和浮点型的数值类 (53), 日期时间 (5) 和一般的对象数据 (5)。一般的对象数据包括会员卡号、性别和工作地点。因此可以认为数据结构化程度较高，而主键会员卡号全部且按照 0-62988 的数据有序排列。存在空值的数据共有如下八项:

表 2: 航空公司数据空值字段及空值率

空值字段	第一次飞行日期	性别	工作地城市	工作地所在省份	工作地所在国家
缺失数	2	3	2269	3248	26
缺失率	0.003%	0.005%	3.602%	**5.157%**	0.041%
空值字段	年龄	第一年总票价	第二年总票价		
缺失数	420	551	138		
缺失率	0.667%	0.875%	0.219%		

从中可以看出，本数据集保存相对完好。最多的数据缺失条数也只有 5% 左右的缺失率。而且缺失数据集中在客户的基本信息方面，这些在后续的价值建模的特征提取部分中仅作为辅助信息。缺失率较高的城市、省份的地理信息在后续的实际分析中仅作为用户整体模型的参考。这些基本信息缺失的原因 (年龄, 性别，工作地点等) 很大可能由于填写录入的数据操作问题，而第一次飞行日期的缺失可能是注册会员在给定时间段内并没有开始第一次飞行。我们主要关注**第一年总票价和第二年总票价**这两个特征，因为其数据的缺失直接影响我们对客户价值模型的分析。

注意到在第一年总票价和第二年总票价这两个特征中，其数据缺失原因不能使用第一年和第二年均没有购买票进行解释，因为在该数据集中存在第一年总票价和第二年总票价数据项为 0 的客户。因此由于其数据缺失率分别仅占原始数据集的 0.875% 和 0.219% 等，全部剔除最多对原始数据集造成约 1% 的影响，可以忽略不计。因此我们直接对原始数据集这两项存在某项为空的用户记录进行剔除。

同时我们需要考虑一些数据异常不满足正常业务逻辑的情况。同样针对第一年总票价和第二年总票价这两个字段，其与总飞行公里数具有对应的关系。即不可能存在两年票价和为 0，但

是折扣率和总飞行公里数同时为 0 的记录。我们对这种客户购票总数和飞行公里数的行为冲突的” 僵尸数据 “ 也直接进行剔除，在两者的基础上，我们只剔除了 944 条数据，保证了数据较高的完整率（98.5%)。进一步地我们也对飞行次数和飞行积分进行了检查，发现在这些数据项中该航空公司的数据均满足业务逻辑。如八季度飞行里程积分加总和里程总积分数据项一致，并且累积飞行次数和八季度飞行次数保持一致等。

至于之前提到的缺失率较高的工作地城市数据，由于总共工作地所在国家数据只有 26 个空值数据，因此我们可以通过工作地所在国家的特征去基本确定会员工作地的大致范围。其中对于 2269 个工作城市的空值数据，对应国家主要: 中国大陆 (1073), 香港 (534), 新加坡 (186), 台湾 (137), 日本 (64) 等。其中只有 25 个数据工作地所在国家为空。进一步发现这 25 个数据工作地 (城市、省份、国家) 全部为空，而另外有一条数据：工作地点省份为尼日利亚，但是国家并没有注明，因此我们对这条数据的工作地所在国家 (NG) 进行补充。

因此我们对本数据集的异常值和缺失值进行了初步的处理和清洗，在处理工作地等文字字符串数据仍然需要进一步的规整，将在下一节中阐述。

3 数据统计分析

在本节中我们对上一节得到的 **62044** 条数据进行进一步的规整化处理，然后对规整化的数据进行相关基本统计信息和结论的刻画。我们按照字段特点将分为**用户特征画像**和**飞行行为画像**的描述。

3.1 用户特征画像

对于用户特征画像的处理主要从用户的地理信息, 身份信息, 在公司的会员信息等角度进行处理和分析。

出于数据规整的考虑，我们首先对会员**所在工作地点**的信息进行部分变换。为方便后续地点标注和统计，我们首先处理**工作地点所在城市**的相关数据。在数据探查后发现，我们发现该字段下存在书写不统一、信息冗余、中英文混杂等情况。首先，该字段下的中文城市混杂了”XX” 和”XX 市” 的两种写法，我们先将这类数据统一为”XX”，即” 哈尔滨市/哈市” 变为” 哈尔滨”,” 乌鲁木齐市/乌市” 变为” 乌鲁木齐” 等。并同时将之前误被处理的其他名字末尾带市的城市恢复。然后，不少工作城市被标注成了”XX 省 XX”，我们同样对这类数据去除了前缀的省份。另外，还有一些中文城市出于种种原因，被标注成了各种各样的英文，我们对这类数据同样进行了处理。在初步的发现中，我们将 ['HK','HONG KONG','HKG'] 等英文单词替换成香港, 将 ['beijing','Beijing','BEIJING'] 等英文单词替换成为北京，另外对哈尔滨, 上海, 大连, 南京等词也进行同样的操作。同时为了进一步补充空值数据，我们对工作地点所在国家为 HK(香港) 和 MO(澳门) 的数据将其所在城市替换为” 香港” 和” 澳门”, 使得数据进一步统一。

经过规整化，我们对工作地点主要出现的城市进行了相关统计。结果如下表所示:

表 3: 航空公司会员工作地点所在城市人数

城市名称	广州	北京	上海	深圳	沈阳	大连	长春	乌鲁木齐	武汉	哈尔滨
会员人数	10315	8184	5127	3870	2393	2068	1904	1805	1348	1257
城市名称	香港	郑州	长沙	佛山	东莞	杭州	珠海	南京	汕头	贵阳
会员人数	1074	831	696	690	574	536	532	515	464	453

我们发现在航空公司国内工作城市所在地的排名中，前 20 的城市大多集中在南方尤其是广东省，如广州、深圳、佛山、东莞、珠海、汕头等城市均有较高的会员人数。另外的城市集中在省会或经济发达的直辖市。而下图中的城市分布同样体现了这一点。这表明该航空公司的主要影响范围在南方尤其是广东一带，这说明该航空公司可能是以经营南方航线为主的公司。

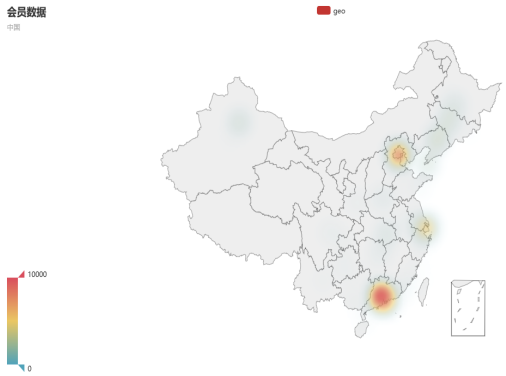

图 3: 航空公司会员分布城市热力图

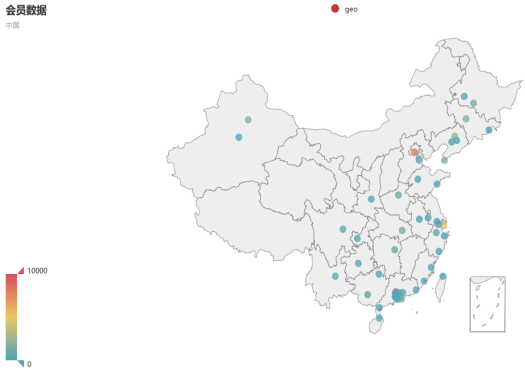

图 4: 航空公司会员分布城市散点图

进一步地，我们按照**工作地所在省份**对中国省份进行了分类，有部分省份被标注成了对应的城市。处于这部分数据量较少的原因，我们不考虑这种情况。而如果工作地点所在国家 (地区) 数据和省市数据出现冲突的时候，我们以工作地所在国家 (地区) 数据为准[1]。当城市数据和省份数据出现不一致的情况，我们以城市数据为准 (如下表中的北京、上海等地)。在此基础上，我们对会员工作地点主要所在省份的数据进行了相关统计。如下表所示:

表 4: 航空公司会员工作地点所在省份人数

城市名称	广东	北京	上海	辽宁	新疆	吉林	湖北	黑龙江	河南	江苏
会员人数	18241	8184	5127	4939	2418	2315	1525	1487	1267	1171
城市名称	浙江	香港	湖南	山东	福建					
会员人数	1126	1018	949	787	739					

从该表中同样可以看出，该航空公司会员主要分布范围是广东，并且会员人数以南方省份为主。但在新疆、东三省等边境往来密切的省份中会员人数也占有较大的比例。下图显示了会员人数按照省/直辖市/自治区分类的分布情况:

[1]这里指工作地点所在如果是港澳台地区但是所在省份并不是其城市的情况

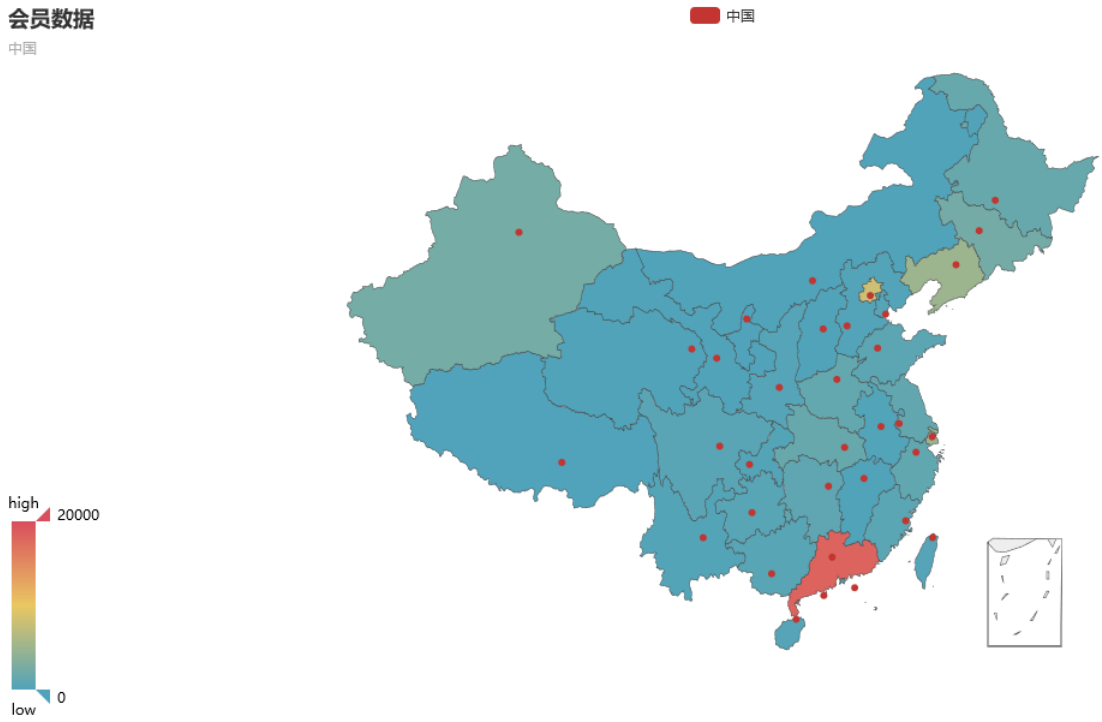

图 4: 航空公司会员分布省份热力图

从中可以清晰地看出，虽然该航空公司在全国各个省份都有会员人数，但是其最主要的会员地点仍然在广东一带，在北京上海两个直辖市。相对来说，通过动态地图的调试我们发现该航空公司会员在南方的人数普遍多于北方，在南方大多数省份会员人数都在 1000 人左右。而在北方较为突出的省份只有东三省和新疆等地区。可以认为该航空公司未来的发展方向是保存并扩大自己在珠三角的用户基础，向自己用户基础较为薄弱的北方推广。

在这一基础上，我们对用户所在国家地区和数据进行分析。虽然该航空公司的主要会员均在中国大陆，但是对于海外的会员仍然不能采用忽略的态度。因为这些会员大多数需要每年进行多次海外长途旅行，需要购买长途机票，因此占据了公司航空收入的较大部分。

我们选择**工作地所在国家**这一数据进行分类，该数据主要由国家的二维代码表示。在初步的数据探查后，我们发现其中也有一些可疑的国家替换项字段和中文名字，如 (’ 中’/’ 北’/’ 沈’/’cn’)。此外为方便 echarts 对于国家的画图，我们引入国家代码与国家英文名的映射数据集 **country_code.csv**。我们同样列出会员人数较多的前 10 个所在国家/地区, 如下表所示:

表 5: 航空公司会员工作地点所在国家/地区人数

城市名称	中国大陆	香港	日本	韩国	美国	新加坡	台湾	澳大利亚	马来西亚	菲律宾
会员人数	56847	986	872	787	568	278	278	271	161	137
会员比例	91.62%	1.59%	1.41%	1.27%	0.92%	0.45%	0.45	%0.44%	0.26%	0.22%

尽管从分类结果看, 该航空公司的会员人数范围非常广阔，总共共约有 110 个国家/地区。但是上表提到这 10 个国家/地区的会员总数已经达到了该航空公司会员总数的 98.5%, 对于分析用

户群体的互项有一定地代表性。从中可以发现，中国内地人数占到了总人数的 90%，剩余一些较多的会员人数主要分布在与中国地理位置相近, 交通贸易较为密切的东亚和东南亚等地，另外还有美国和澳大利亚华人较多的国家地区。下图则展示了世界范围内海外[2]不同地区的会员人数分布图:

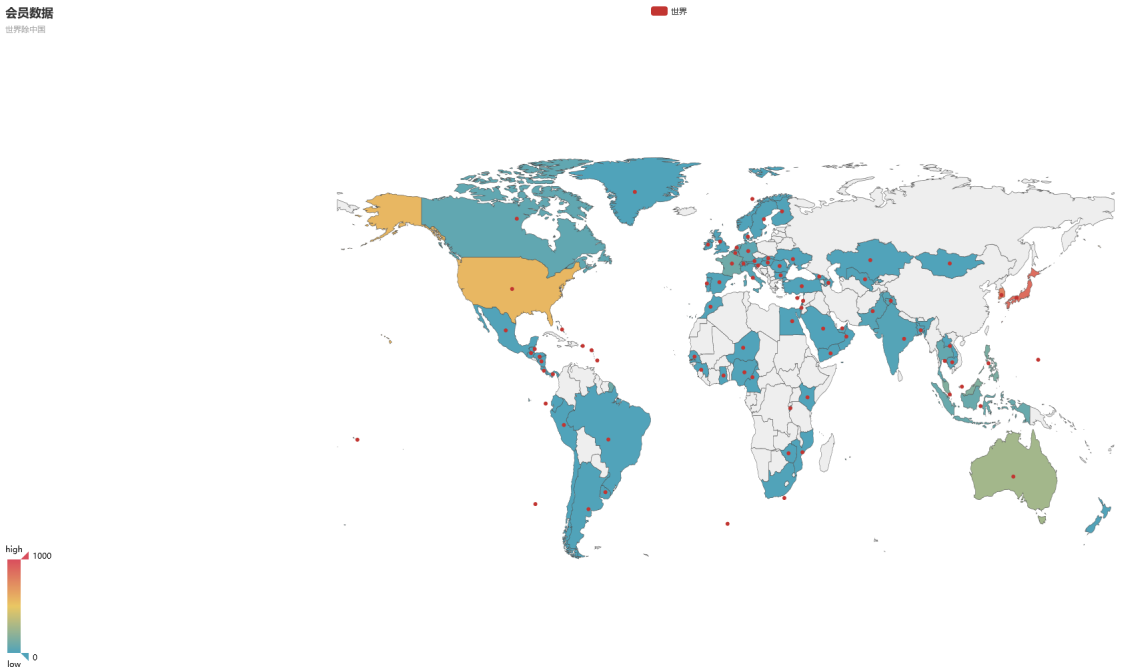

图 5: 航空公司会员分布海外国家/地区热力图

可以看到该航空公司的会员人数分布在世界各大洲，但是主要集中范围仍然是东亚、东南亚等地，这与该航空公司拥有最多会员地点的广东 (广州) 也有一定关系。广东与东南亚各地文化人员等交流较为密切，因此该航空公司因此开拓了较多的东南亚和东亚航线，以此更好地满足客户的需求。而整体上，面对海外各国的不同需求，航空公司仍然需要积极完善并发展海外航线的服务，扩大其用户群体的差异性以应对区域性风险。

在地理信息的基础上，我们希望对客户的基本身份信息有一个更清晰的描绘，我们首先对该航空公司会员中不同等级和**性别**的关系。在**会员等级**共有 4, 5, 6 三个级别的前提下, 该航空公司主要的客户群体仍然是男性且会员等级为 4 级的会员。具体比例如下图所示:

[2]我们在世界地图范围去除了中国的数据更好地展示不同国家/地区的差异

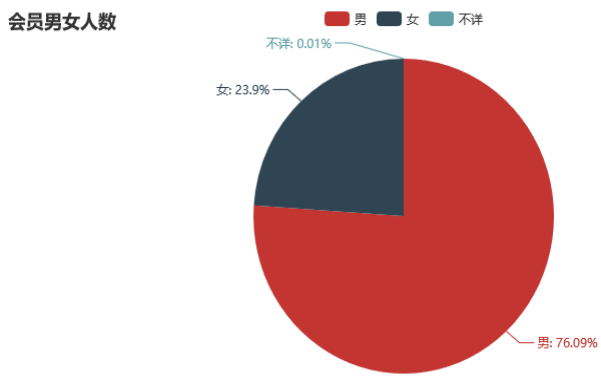

图 7: 航空公司会员性别分布图

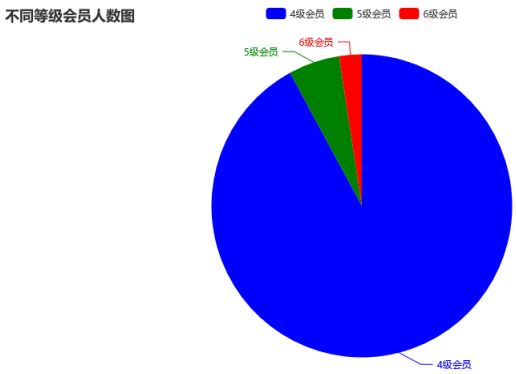

图 8: 航空公司会员会员等级图

目前该航空公司超过 92% 的会员仍然是等级最低的 4 级会员。因此在保证会员整体基数扩大的基础上，航空公司更应该想办法关注会员等级的转化与升高，发挥用户的发展潜力和利润空间，这样才能让航空公司获取到更大的利润。

年龄的分类也是航空公司推测不同会员可能购票行为背后的原因。比如年龄较小的会员进行购票行为可能出于出国旅游、留学等因素，一般来说青壮年的会员进行购票飞行行为可能更多地出于平时的商务活动交流和工作需要；而年龄较大的会员则乘飞机的原因可能在于旅游。针对不同年龄段的客户，航空公司也因此制定不同策略进行营销。而该航空公司会员用户年龄段分布较广，在不考虑空值数据的情况下, 会员最小年龄的会员仅为 6 岁，最大年龄的会员则有 110 岁。平均会员年龄为 42.5 岁。我们对会员的年龄数据这一可近似看作连续类型的数据进行了 10 等分的等宽离散化和 5 等分的等频离散化的操作, 发现绝大多数用户 (超过 60%) 处在 34-50 岁之间。同时，在青壮年的数据中该航空公司的会员人数以男性为主：

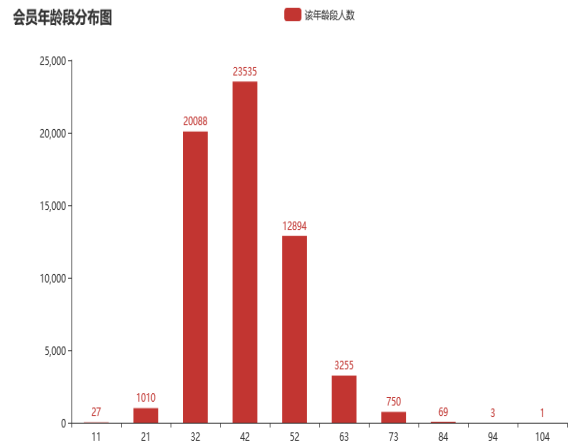

图 9: 航空公司会员年龄分布图

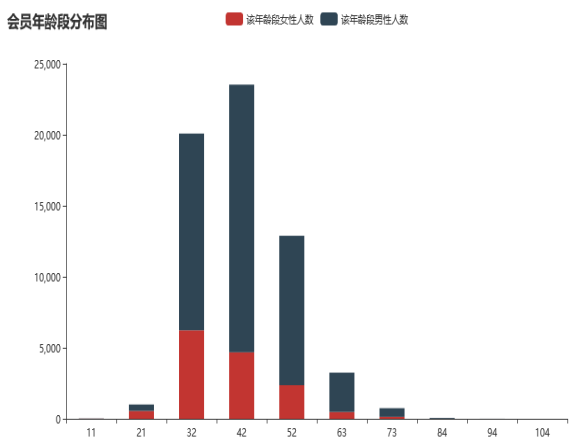

图 10: 航空公司会员年龄 × 性别等级图

针对青壮年男性可能因为经常工作商务出差选择成为会员的情况，该航空公司可适当进行品质服务的推广和差异化定价。这部分人群往往可能临时出差往往在飞机起飞最后几天才购票，并且获得工作单位的报销，因而往往对高票价不敏感, 因此往往能接受较高票价和较高等级的舱

位。航空公司可以选择增强自身的服务水平，提高这部分用户的满意度和忠诚度。

3.2 用户飞行行为画像

在对用户基本身份信息的分析后，我们着重对用户飞行模式进行总体上的分析。我们重点关注飞行次数、飞行时间和入会时间的对比、不同积分之间积分相关性等特征。

关于会员的**飞行次数**，在基本统计的结果后发现所有会员的平均飞行次数为 11.97 次，最少的会员飞行次数为 2 次。而最多的会员则在八个季度总共飞行次数达到 213 次。在对客户飞行次数进行等频离散化后，发现 80% 的会员飞行次数主要分布在 18 次及以下。其中 26% 的会员飞行次数只有 2-3 次左右。而具体的等宽离散化后客户飞行次数段结果如下, 可见大多数会员仍然飞行次数在一个月一次以下。

表 6: 航空公司会员不同次数飞行段人数

序号	1	2	3	4	5	6
飞行次数段	(2,23]	(23,44]	(44,65]	(65, 86]	(86, 107]	(107, 128]
会员人数	52866	5843	1549	521	167	67
会员比例	**85.207%**	9.418%	2.497%	0.840%	0.269%	0.108%
序号	7	8	9	10		
飞行次数段	(128,149]	(149,170]	(170, 191]	(191, 213]		
会员人数	19	5	3	4		
会员比例	0.031	%0.008%	0.005%	0.006%		

同时针对不同飞行次数段的男女人数对比，我们有如下的结果。可以看出在低飞行次数的比较中，男女性的比例和总体比例相当，甚至女性飞行比例略高于男性。而随着飞行次数的增加，男性的比例均保持在 80% 以上的水平。因此可以认为，大多数飞行次数较多的会员为男性。

图 10: 航空公司会员不同次数飞行段男女人数

此外，我们想要比较**会员的入会时间和第一次飞行时间的差异变化**，以分析在 1998-2007 年这段时间航空公司客户量变化和飞行的趋势。出于篇幅考虑我们仅分析对年和月分别进行处理。结果如下图:

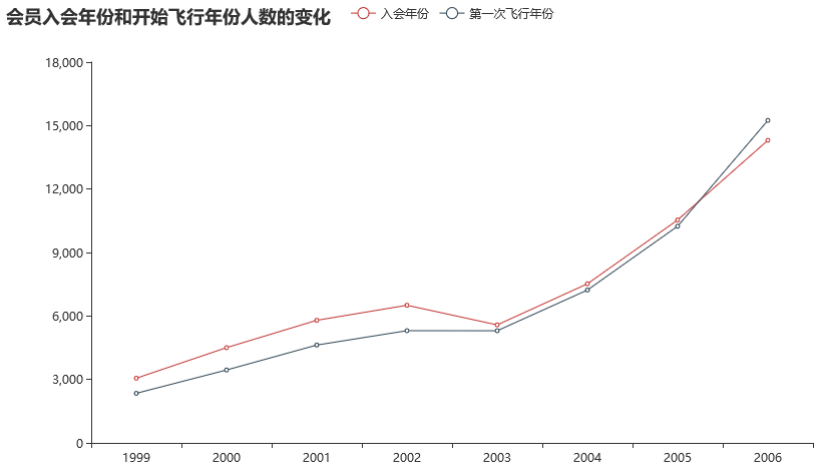

图 11: 会员入会年份和开始飞行年份人数对比图

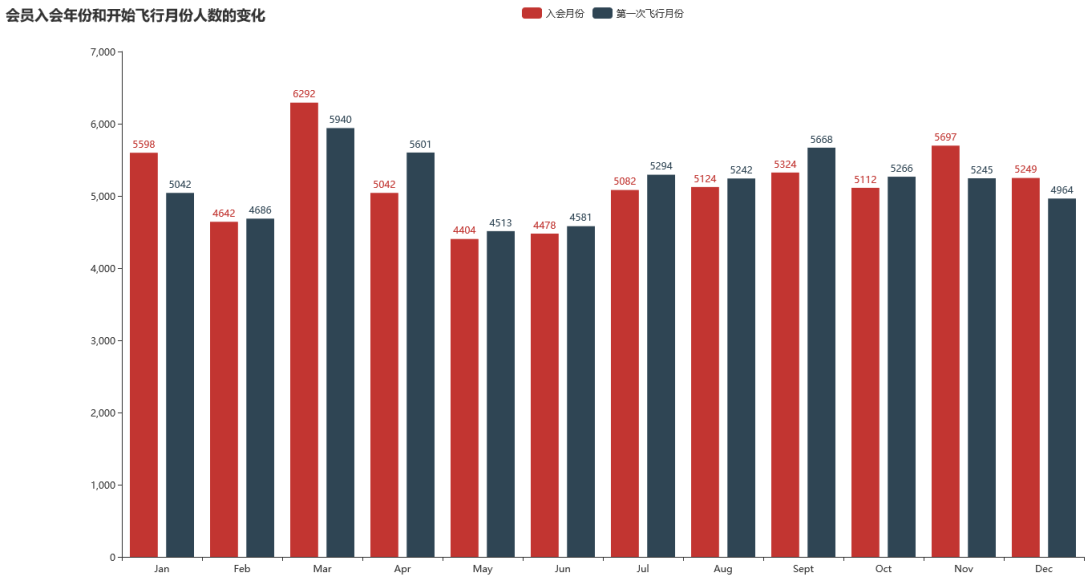

图 12: 会员入会时间和开始飞行月份人数对比图

我们对这两个字段均截取了 1999-2006 年八月的数据，结果显示在这八年按年度变化来看，会员入会时间和开始飞行时间以 2003 年为节点前后有一个突破性的增长。之前变化趋于平缓甚至增长人数减慢，而之后变化则增长较为迅猛。说明航空公司在 2003 年迎来了较好的发展机会使得其入会人数和飞行人数实现了一个较大的突破。而从月份来看，对于该航空公司来说，并没有出现较为明显的淡季和旺季的区别。在不同的月份中入会人数和第一次飞行人数这两个量均在 4400-6300 的范围内波动，且不同月份之间差别不是很大。

而具体到两者的差值，从整个数据平均来看，会员在注册会员之后约 180 天内便开始第一

次飞行，有超过 80% 的会员选择在注册会员的第一年开始第一次飞行。但是不乏有注册会员超过八年才开始飞行的客户。

而针对不同等级的会员，其在**观测窗口的飞行次数和基本积分**存在系统性差异。直观上看在八季度的观测窗口中飞行次数和基本积分的变化可以反映会员在航空公司的参与, 流失和对航空公司信任的趋势。以下是平均飞行次数和按照不同会员等级八季度飞行次数的变化关系:

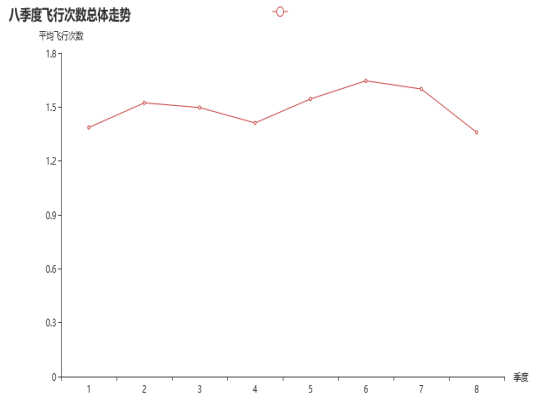

图 14: 航空公司会员观测窗口飞行次数变化图

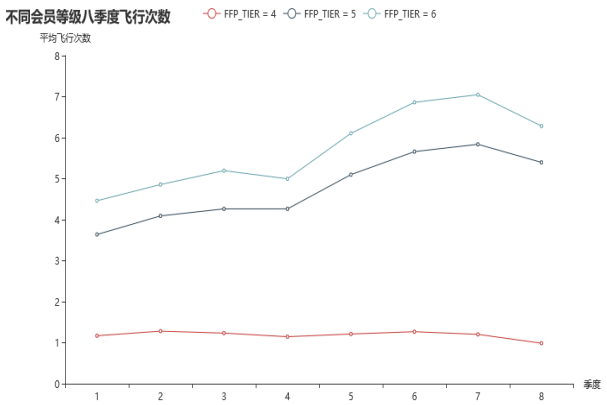

图 15: 航空公司会员观测窗口飞行次数变化图(会员等级)

总体上来看，航空公司会员每季度观测窗口的飞行次数平均在 1.5 次左右。在 2-3, 6-7 季度有个略微明显的上升。而在不同会员等级的飞行次数存在明显差别。随着会员等级的升高，这些会员平均每季度的飞行次数明显上升。平均来看，等级为四级的会员虽然占据了会员人数的绝大多数，但是每季度的飞行次数仅有 1.2 次左右。而另外两个会员等级的会员随着季度的推移，飞行次数总体上保持增长的趋势。这点同样可以从下图的积分变化图看出，基本积分反映了会员参与航空飞行的积极程度，平均来看会员每季度积分的贡献程度约为 1200-1300 积分，而其贡献度最大的仍然是会员等级最高的用户，他们往往能每季度贡献约 6000-8000 积分且保持了较高的增长趋势。这说明会员等级与飞行次数和飞行次数等用户行为具有较强的相关性。

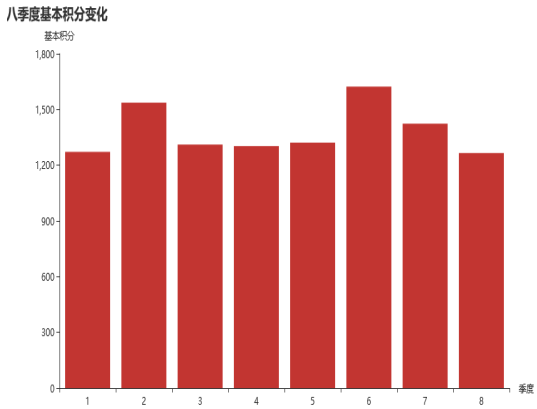

图 16: 航空公司会员观测窗口基本积分变化图

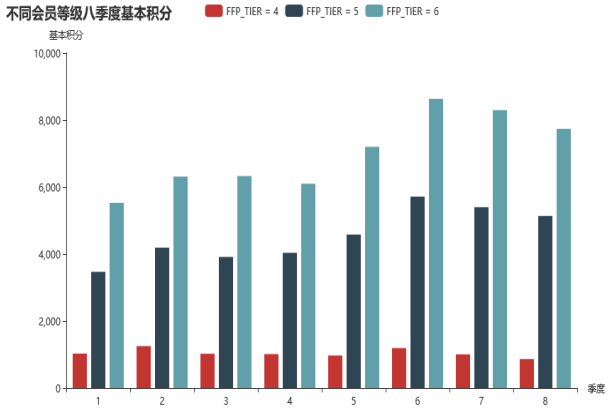

图 17: 航空公司会员观测窗口基本积分变化图(会员等级)

而具体到用户的行为和主要观测变量 **(飞行次数，票价，非乘机积分总和)** 等，我们通过相关关系图和统计指标表[3]进行分析:

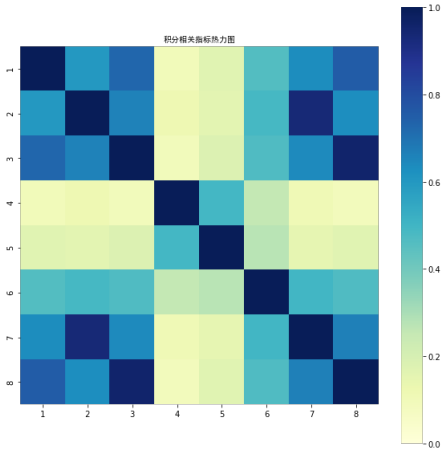

图 17: 用户相关行为热力图

字段序号	字段名称	均值	变异系数
1	精英资格积分 (2yr)	269.73	6.15
2	总票价 (1yr)	5389.30	1.51
3	总票价 (2yr)	5676.83	1.54
4	其他积分 (1yr)	544.70	7.30
5	其他积分 (2yr)	822.63	6.26
6	积分兑换次数	0.32	3.53
7	里程积分 (1yr)	5423.4	1.58
8	里程积分 (1yr)	5634.37	1.67

图 18: 用户相关行为统计指标表

从中可以看出在用户的这些积分指标中不同用户之间差异较大，变异系数 (CV) 很大都超过了 3。而用户的积分兑换次数与总票价和里程积分均有较强的相关性，超过了 0.6。用户的这些积分指标之间均具有较强的正相关性。说明用户在这些积分的表现和乘机行为、在该航空公司的活跃程度具有较高的相关性

同样在这些指标上，会员等级也有较大的分层差异。更高等级的用户在这些指标中均有更大的积分和票价等。具体可以见下表所示:

表 7: 航空公司不同会员等级平均积分指标表

会员等级	1	2	3	4	5	6	7	8
4	32.25	4321.68	4140.39	450.08	630.03	0.21	4316.91	4033.67
5	1765.46	15414.61	20597.10	1334.18	2513.15	1.37	15612.32	20830.22
6	5873.15	23132.61	30100.51	2341.17	4289.83	2.17	24267.26	31865.86

倘若航空公司能够用相关政策刺激会员对其他积分的获取，对级别较低的会员也能起到一定的激励作用并能够在其他积分的使用中转变为对票价和里程的贡献，使得会员等级更偏向于较高等级, 进而提高航空公司的整体收入。

[3]去除第一年精英资格积分这一变量是因为其所有数据均为 0, 不具有参考意义; 其他积分主要指合作伙伴, 促销, 外航转入等情况

4 数据挖掘分析

4.1 客户端与企业端心智模型的交互

在航空公司对于客户关系管理的过程中，其核心是**实现顾客价值和企业收益的最大化双方的平衡**。对于企业来说，实现高质量的价格歧视能够为更加精准地获取更多的生产者剩余，同时又可以使顾客能够获得相对正的效用。在此基础上，航空公司管理者心智模型的便是其对顾客行为的认知和评价。在前面基本数据的统计中，航空公司认为在其会员群体中，不同等级的会员在价值等多方面上差异较大。因此需要针对不同用户制定相应的价格与服务策略，在优先发现并识别高等级会员价值的前提下, 尽最大可能满足不同等级的会员的需要, 因此能够对不同会员进行制定策略。在建立客户关系时，企业应当把重点精力放在他们认为价值较高的客户群体上，提高他们的忠诚度，并吸引其他类型的用户向这类用户转变。

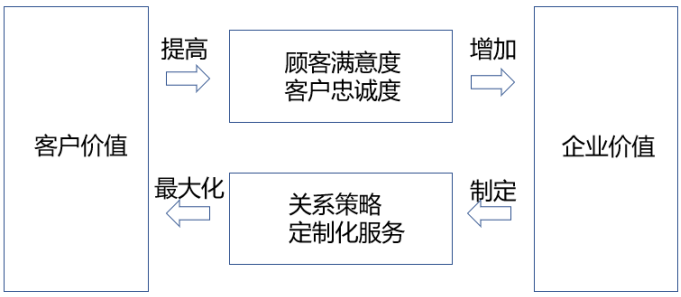

图 19: 客户与企业心智模型的转化

同时从数据来看, 航空公司管理者认为该航空公司的吸引力主要集中在中国南方 (尤其是广东) 和北方的大城市等，在中小城市的市场仍然有较大潜力。该航空公司的用户群体仍然以青壮年为主，这部分人群拥有相对较大的发展潜力和利润空间。

4.2 客户价值识别模型

进一步地为了更好地识别和发现客户的价值，航空公司需要通过客户数据的指标进行不同价值的识别。在客户价值识别应用中目前最广泛的是 RFM 模型，分别表示最近消费时间的间隔 (**R**ecency), 消费频率 (**F**requency) 和消费金额 (**M**) 对用户进行细分。在此基础上公司能够较为精确的判断用户的价值，从而更好地为公司的营销决策提供给参考和支持 ([1]。

而具体到航空公司的案例中，我们仿照 ([3],[4]) 数据挖掘的做法。由于数据集的全面，航空公司客户的价值可能会受更多因素影响。以与航空公司利润最为密切的航空票价为例，其明显地受到飞行距离, 舱位等级等因素的影响。单次支付相同航空票价的用户对航空公司整体的利润影响有很大差别。因此这里我们将原来的消费金额这一指标细分为客户一定时间内累积的飞行里程 (M) 和乘坐舱位对应的平均折扣率 (C) 两个指标。此外，会员的入会时间长度对客户价值的维持也具有较大的影响。因此我们在模型中增加客户关系长度 (L) 这一指标，进而更加细分不同客户群体的价值。

因此我们将上述提到的指标作为航空公司识别客户价值的指标，并将其记为 **LRFMC 模型**。

表 8: 航空公司客户价值识别模型各指标含义

模型字母	含义	相关字段
L	会员在入会时距观测期结束的时间	LOAD_TIME - FFP_TIME
R	会员在最近一次乘机距观测期结束的时间	DAYS_FROM_LAST_TO_END
F	会员在观测期内累计乘机次数	FLIGHT_COUNT
M	会员在观测期内累计飞行里程	SEG_KM_SUM
C	会员在观测期内平均折扣率	avg_discount

对这五个指标进行数据提取之后，发现我们提取的这五个指标单位差异较大。而采用分割聚类 (partition clustering) 这一方法在计算会员数据点距离的时候对数据指标的大小较为敏感，因此我们需要对数据进行标准化处理。这里我们采用的方法为 **Z-Score 方法**，也即将指标通过均值和标准差进行标准化, 并将其标签修改为 [ZL, ZR, ZF, ZM, ZC]:

$$Z_i = \frac{x_i - \mu}{\sigma}$$

这样得到的数据基本统计信息如下表所示:

表 9: LRFMC 模型各指标取值范围

属性名称	L	R	F	M	C
平均值	0.00	0.00	0.00	0.00	0.00
标准差	1.00	1.00	1.00	1.00	1.00
最小值	-1.32	-0.94	-0.71	-0.81	-3.17
最大值	2.30	3.08	14.25	26.76	4.21

在此基础上，我们根据这五个指标的数据对客户进行聚类分组。然后结合业务知识对各个客户群进行特征分析，分析其客户价值。我们预先考虑可能从 2 到 9 的聚类数目，并计算其相应的样本内平均误差 (SSE) 如下图所示。由于在聚类数目从 5 个变化到 6 个的时候, SSE 数值变化差距较小。因此我们认为对本数据集选取**聚类数目为 5 个**比较合适。

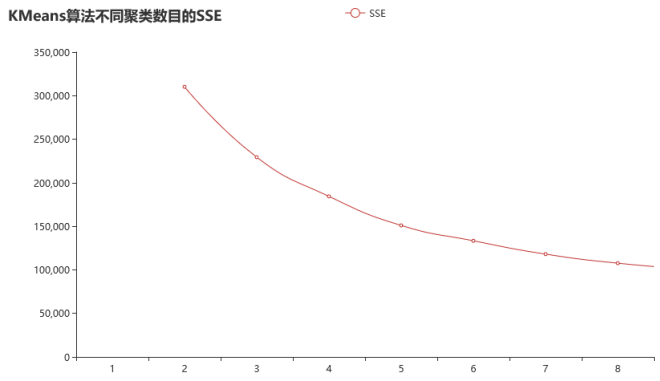

图 20: LRFMC 模型 KMeans 算法不同聚类数据对应的 SSE

然后对此数据进行聚类分组的分类数目和各个簇聚类中心的结果如下表所示:

表 10: LRFMC 模型总体客户聚类个数与中心结果

聚类类别	聚类个数	ZL	ZR	ZF	ZM	ZC
A 类用户	24626	-0.700	-0.415	-0.161	-0.161	-0.257
B 类用户	15738	1.161	-0.377	-0.087	-0.094	-0.157
C 类用户	5337	0.483	-0.799	2.483	2.424	0.309
D 类用户	12117	-0.313	1.687	-0.574	-0.537	-0.175
E 类用户	4226	0.044	-0.003	-0.230	-0.235	2.175

并且我们简要地列出不同用户的人数特征分析见表11所示。我们主要判断不同客户群在 L/R/F/M/C 属性的大小关系。这样能够得到比如 F, M, R 是 C 类用户的优势特征, L, C 是 A 类用户的劣势特征等，从而我们能够具体从每个客户群的特征中观察其规律。

表 11: LRFMC 模型客户群特征描述表

聚类类别	优势特征			劣势特征		
A 类用户				**L**	**C**	
B 类用户	**L**	F	M			
C 类用户	**F**	**M**	**R**			
D 类用户				**F**	**M**	R
E 类用户	**C**			**R**	F	M

Notes. 其中黑体加粗的部分表示其是最大值或最小值

因此从中可以看出每个客户群具有显著不同的表现特征。直观上可从如下的图中进一步分析, 可以看出不同类客户之间在 LRFMC 特征上有着较大差距：

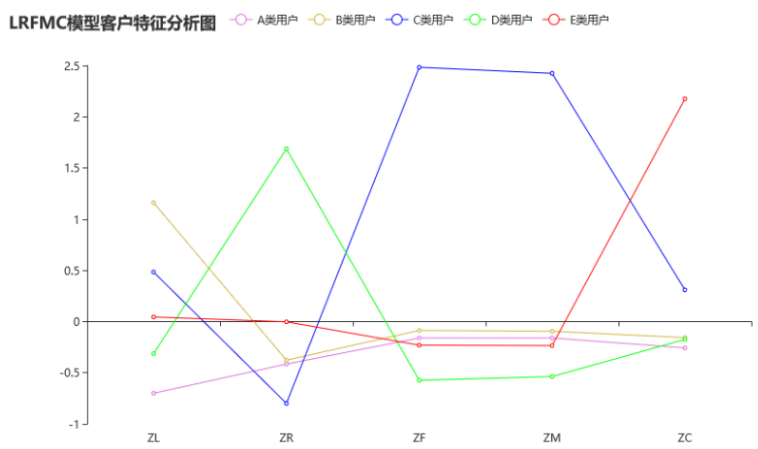

图 22: 航空公司会员客户价值分类图 A

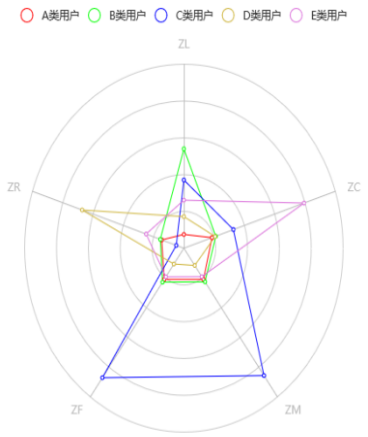

图 23: 航空公司会员客户价值分类图 B

然后我们仿照[4]的结果我们定义五个不同的客户级别:

重要保持会员: 这类客户经常乘机并且选择航班较高等级的舱位，对应航空公司的高价值客户。航空公司应当将最优质的资源投放到他们身上，进行针对性营销，保证这类会员的忠诚度和满意度。

重要发展会员: 这类客户折扣率一般，最近经常乘机，但是其入会时间短累计里程和次数较少。对应航空公司的潜力客户。航空公司应当争取这类客户能够经常在本公司乘机消费，提升这类的满意度。

重要挽留会员: 这类客户的乘坐次数和里程较高，但是很久没有乘机。公司应当了解其最新动态，尽可能努力挽留这类用户。

一般会员: 这类客户乘机频率较低，入会时间较短，发展潜力不大。公司应采用正常宣传发展的态度，希望尽可能使其变为高价值的会员。

低价值会员: 这类客户乘机频率较低，入会时间不短，折扣率较低。现有价值和价值潜力不大且转换难度较大，公司可以不用管这类用户。

这些分类的差异进一步也可以在这些会员八个季度的飞行次数和积分情况图中看到，可以看到 D 类用户价值较低，在后面几个阶段甚至没有乘机记录和基本积分；相反的 C 类用户保持了较高的活跃度和乘坐飞机的行为, 并且作为航空公司的主要利润来源，飞行次数远远高于其他类别的用户。

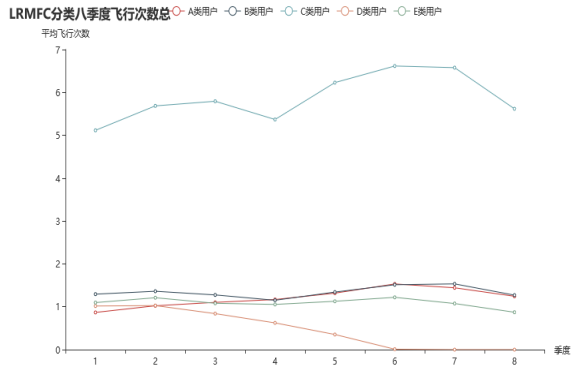

图 24: 航空公司会员客户价值分类与飞行次数图

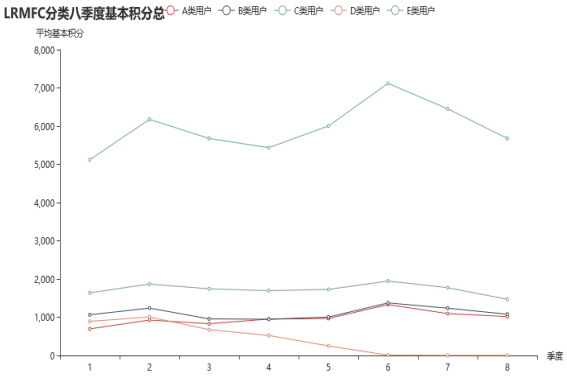

图 25: 航空公司会员客户价值分类与基本积分图

结合上面我们对五个客户级别的定义我们认为不同群体分别对应如下表格中的客户，而下图可以看出不同会员等级与其相应的 LRMFC 模型占比不同关系, 其中会员等级高于四级的用户中价值较低的 A 和 D 类用户较少，而对应六级的会员拥有占比最高的 C 和 E 类用户。因此可看出因此航空公司这一会员**等级分类的机制设计**对于顾客价值识别起到了逆向选择和信号识别的效果。

客户分类	排名	对应群体
C 类用户	1	重要保持会员
E 类用户	2	重要发展会员
B 类用户	3	重要挽留会员
A 类用户	4	一般会员
D 类用户	5	低价值会员

图 25: 用户 LRFMC 模型与客户实际群体对应关系表

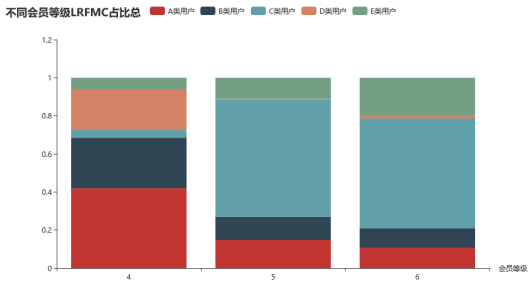

图 26: 用户等级分类与其 LRMFC 占比图

而这个 LRFMC 模型可以被推广至更多层次的细化模型。比如而在不同年龄段/不同性别之间各个用户份额的占比也有明显的区别 (见附录7)。

上述的聚类过程已经将客户分成了五个等级。由于所选数据集样本量较大，因此该模型的鲁棒性较好。而基于这些历史数据我们可以对未来客户的价值进行分类判别。我们将 LRMFC 模型的五个基准特征和用户其他主要特征结合，试图分析其对分类模型最终结果的重要性影响。

我们对对训练集和测试集随机划分为 80% 和 20% 的部分。主要采用的主流机器学习模型有随机森林 (RF), 多层感知机 (MLP), 梯度提升树 (GBDT), XGBoost，实验设置采用 sklearn.emsembler.RandomForestClassiifier, sklearn.emsemble.GradientBoostingClassifier, sklearn.neural_network.MLPClassifier, xgboost.XGBClassifier 这些分类器的默认设置。这些基准模型在训练集和测试集的结果如下表所示：

表 12: 经典机器学习在预测 LRFMC 模型的准确率表现

模型选择	随机森林	梯度提升树	多层感知机	XGBoost
训练集	**99.93%**	99.46%	99.84%	99.06%
测试集	98.36%	98.76%	**99.48%**	98.45%

从中可以看到经典的机器学习模型对原来模型起到了较好的预测效果，而在这些指标中，对结果影响较大的除了本身用户聚类的指标:L/R/F/M/C, 这五个因素的重要性累计占到了总体分类指标的大约 90% 左右, 此外一些比较重要的特征有**总基本积分，平均乘机时间间隔，飞行次数，单位里程票价**等标签。

尽管不同模型之间略有差异。但是这些普遍在测试集上取得了超过 **98%** 的准确率。这说明现有数据对于判断客户价值已经达到了非常准确的精度。因此在实际业务中, 航空公司可以定期运行一次平台数据集累计用户的聚类, 根据其与不同聚类中心的距离对其进行分类。同时对新增用户进行判断，并观测其用所标签定义的用户的实际发展情况。如果模型与实际结果差异较大，则需要重新选择标签对模型进行重新训练。而针对建模后不同价值类用户，下文则会具体阐述对应分类用户的管理和营销策略。

4.3 客户流失模型

而在现有的 LRFMC 模型对于客户消费行为模型的构建中，并没有对老客户的流失情况做具体的分析。许多航空公司虽然名义上有很多会员，但是很多在经历开始消费的初期热潮之后往往不再进行消费。在激烈的航空公司之间的竞争中，我们应着力关注老用户的流失。统计结果表明，为了弥补失去一个老客户的损失，公司争取到一个新客户的成本大约是留住一个老客户成本的六倍 ([3])。

在当前国内市场竞争日益激烈的今天，航空公司在客户流失方面应当引起较大的重视。如何改善流失问题并因此提高客户忠诚度，是航空公司进一步维持自身市场, 实现会员可持续性发展的重要举措。

鉴于此，我们将建立航空公司老客户的**会员流失分类模型**。我们将观测期飞行次数大于 6 次的客户定义为老客户，并给出如下表内关于流失的定义和用户数据所占的比例 ([3],[4])：

表 13: 客户流失模型不同老客户的分类情况

分类序号	用户名称	用户定义	用户比例
0	未流失客户	第二年飞行次数与第一年飞行次数比例大于 90%	57.39%
1	准流失客户	第二年飞行次数与第一年飞行次数比例在 [50%, 90%)	22.27%
2	已流失客户	第二年飞行次数与第一年飞行次数比例小于 50%	20.34%

同时我们选取客户信息中的相关属性进行。我们随机选取数据的 80% 作为训练样本，剩下的 20% 作为测试样本，从而构建客户的流失模型。希望能够运用该模型预测未来客户的类别归

属失。在本节中我们仍然采用之前提到的机器学习模型对数据进行标准化，然后将这些特征带入模型进行学习。

首先, 出于模型分类的考虑且更加直观反映结果，我们首先将未流失客户和准流失客户看作一类进行讨论。具体结果如下表所示:

表 14: 经典机器学习在预测客户流失模型二分类的准确率表现

模型选择	随机森林	梯度提升树	多层感知机	XGBoost
训练集	**98.58%**	86.85%	87.31%	86.45%
测试集	85.92%	**87.01%**	86.17%	86.98%

可以看出，不同机器学习模型在测试集的数据表现差异不大，大约能够达到 85%-87% 的准确率。表现相比之前正样本已经达到 79.66% 来说，提升不够显著。这可能是因为对已流失客户的数据样本学习偏少，对其特征的识别不够到位。

在此基础上，我们为了进一步衡量该分类模型的准确度使用模糊矩阵和 ROC 曲线进行判别，出于篇幅考虑，我们在下图仅列出随机森林和多层感知机的分类模型模糊矩阵和 ROC 曲线进行分析：

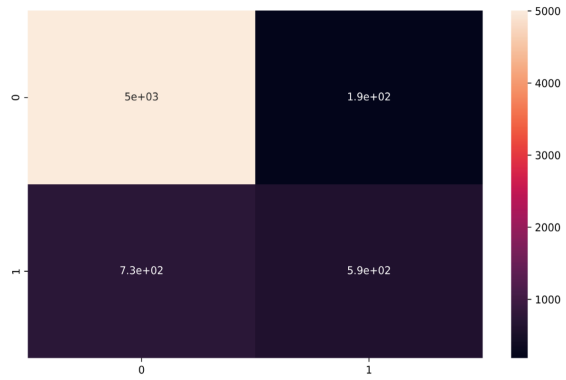

图 28: 随机森林二分类模糊矩阵

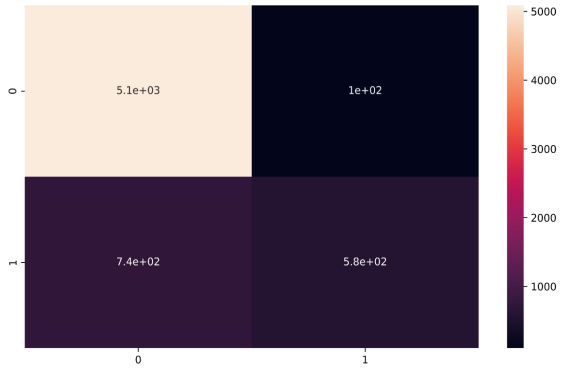

图 29: 多层感知机二分类模糊矩阵

通过计算，我们发现随机森林模型的命中率 (真正类率)$\frac{TP}{TP+FN}$ = 87.3%, 特异性 (真负类率)$\frac{TN}{TN+FP}$ = 75.6%。而从这个尺度上看多层感知机的表现更好，它的命中率达到了 88.9%, 并且特异性也达到了 85.3%。而从下面的 ROC 曲线 (关于真阳率和假阳率) 的对比中也可以看出，多层感知机的 ROC 曲线面积会比随机森林分类的面积稍大一些，这也说明多层感知机的模型表现相对较好。

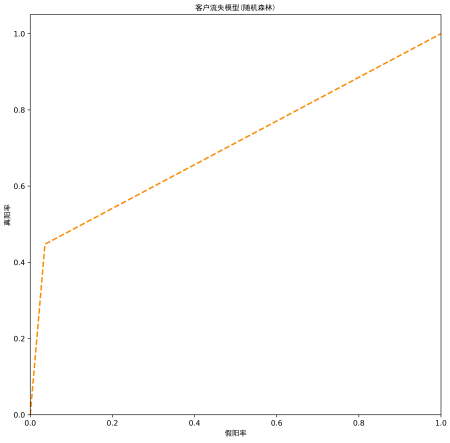

图 30: 随机森林二分类 ROC 曲线

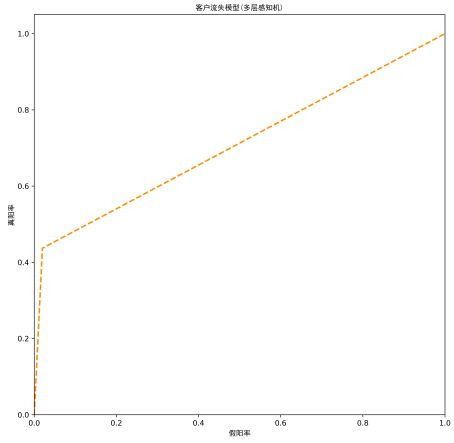

图 31: 多层感知机二分类 ROC 曲线

而倘若我们使用原始的三分类模型 (未流失用户, 准流失用户, 已流失用户) 去判断识别客户流失模型，预测结果可见下表所示:

表 15: 经典机器学习在预测客户流失模型二分类的准确率表现

模型选择	随机森林	梯度提升树	多层感知机	XGBoost
训练集	**98.17%**	68.65%	68.61%	67.31%
测试集	64.48%	67.23%	**67.71%**	67.17%

可以看出，不同机器学习模型在测试集的数据表现差异不大，大约能够达到 64%-68%。表现相比之前正样本已经达到 57.4% 来说，提升不够显著。这可能是因为对准流失和已流失客户的数据样本学习偏少，对其特征的识别不够到位。比如在下文梯度提升树和 XGBoost 的混淆矩阵可以看出, 其识别正负样例的准确率相较之前均比较低 (**(梯度提升树)** 未流失:68.36%; 准流失:44.44%; 已流失:70.59% **(XGBoost)** 未流失:67.04%; 准流失:41.74%; 已流失:73.40%)。在这种情况，航空公司可能需要进一步地采样从而平衡正负样例进而获得更好的效果。当然对中间准流失客户边界的定义可能会导致不同模型在准流失用户的识别中判断较为不准确，这可能航空公司需要进一步根据业务需求定义不同流失用户。

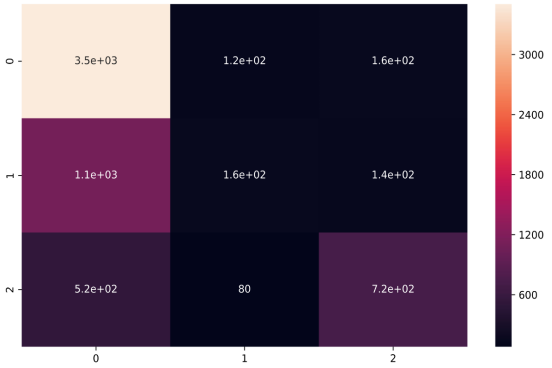

图 32: 梯度提升树三分类模糊矩阵

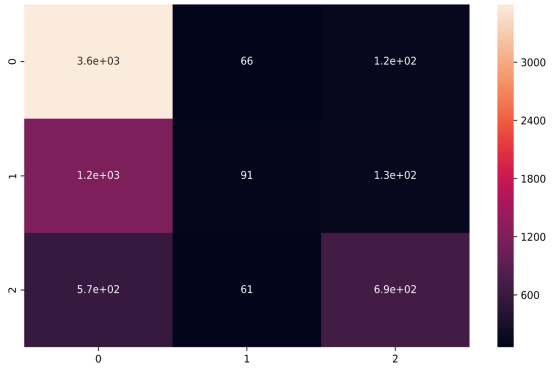

图 33: XGBoost 三分类模糊矩阵

而不管在二分类和三分类模型中，我们发现部分因素始终在识别客户流失模型中扮演着关键作用。下图显示了 XGBoost 在识别模型标签中起到较为重要作用的特征, 其与随机森林和梯度提升树给出的重要特征基本保持一致。**最后一次乘机的时间，乘机频率，乘飞机的（最大/平均）间隔，积分变动次数，总精英积分**等标签被认为是反映用户在航空公司活跃度特征的内容, 对于衡量客户流失模型起着较为重要的作用。航空公司可根据这些重要的特征进行主成分分析，得到更加直观且稳健的结论。

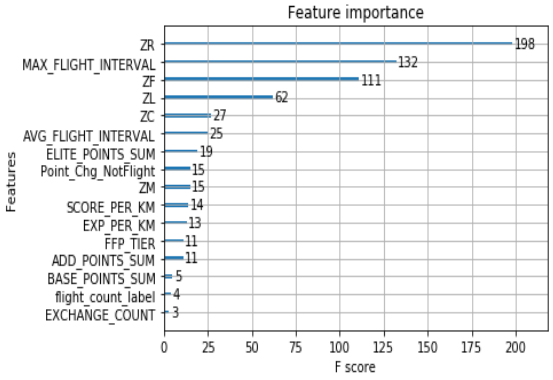

图 34: XGBoost 二分类重要特征

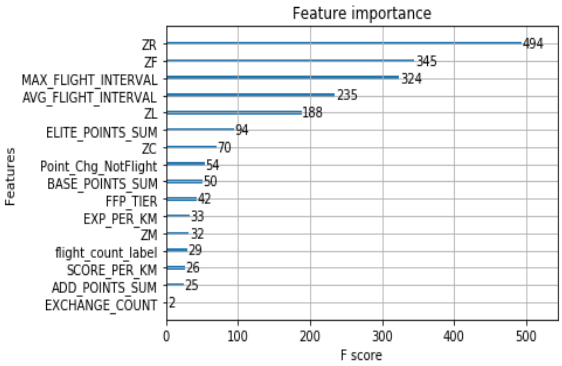

图 35: XGBoost 三分类重要特征

在上面中我们构建了基于乘飞机次数指标评价的客户流失模型，航空公司可以因此针对不同的老客户制定相应合理的营销策略。比如针对准流失客户可以针对性地营销尽可能留住这些老客户，比如对他们提，增强他们对航空公司的忠诚度，从而能够延长客户在航空公司中的消费周期和时间 ([3])。

5 商务分析与应用前景

在之前的客户价值识别模型和客户流失模型中，我们分别针对客户的不同行为构建了相关指标。而这与商务分析中的产品生命周期法对应的不同阶段也紧密相关。航空公司面对的不同会员可以看作是产品的不同阶段, 这也正好对应企业的不同策略 ([5])。如下表所示:

表 16: 客户价值建模与产品生命周期法的联系

模型选择	研发期	引入期	成长期	成熟期	衰退期
客户		重要发展会员	重要保持会员	重要挽留/准流失会员	已流失/低价值会员
利润率		较高	高	下降	低
风险	高	高	较高	较低	低
竞争/流失率	极少	少	增加	最多	较多

面对不同的客户群体，航空公司应当制定相应的策略。比如如下的措施可以参考 ([3], [4]):

会员升级提醒: 我们看到该航空公司拥有三种等级会员, 不同等级的会员可以享受航空公司不同的服务, 如候机室和商务舱优惠等。而达到更高级的会员需要满足一定的飞行里程和飞行次数，同时倘若客户活跃度降低，公司会根据客户最近的记录重新评价客户的会员级别。为了鼓励并激励会员更加地参与公司会员升级的互动，航空公司可以对快要达到要求的会员进行提醒和展开一些促销活动，以提高会员的参与度。倘若升级成功，自然能够提高客户对公司的忠诚度。

联合销售: 航空公司可以通过与其他非航空类企业合作，通过对于用户信息的识别, 进行相关用户可能感兴趣的产品的销售。不同用户乘机行为本身受工作地点, 时间等原因的限制，飞行次数通常可能有一定的上限。但是客户在与其他非航空类合作处的积累积分，同样可以使用到该航空公司中去，这样从企业与企业的合作变为到企业与客户的合作。

积分兑换提醒与优惠: 我们看到不同群体的会员之间在基本积分, 里程积分, 精英积分等差异较大。而这些积分与客户的里程数具有紧密的正相关性。当客户的里程和飞行次数达到一定程度时候，可以选择兑换。但是客户对这类信息关注敏感度较低，而白白浪费很多机会。同样地，航空公司可以向即将达到兑换要求的会员发出提醒，这样增加了会员进行自我流失的成本，从而提高客户对公司的满意程度.

而具体到上述不同生命周期法的阶段，结合之前两部分对于客户的分类和相应采取的措施来看, 企业要获取长期的利润就必须要在选择不同周期会员的优先级的前提下，尽可能保证更加稳定且有质量的客户关系。从 LRFMC 的客户价值模型和客户流失模型中, 在引入期和成长期的用户是航空公司最需要保证的利润来源，航空公司针对这类会员需要最高级别的服务水平，因此才能提高并保证自己在客户心目中的形象。而倘若用户经过成熟期之后最近的里程和飞行次数下降，航空公司应当尽力了解这类客户可能流失的原因并尽力保证这类客户的用户基础, 延长这类会员在该航空公司的消费时间。

6 总结与建议

本文对航空公司数据进行了初步分析和数据挖掘。首先对提供的客户信息进行了相关整理。首先我们按照业务逻辑和常识对数据进行了异常值处理和缺失值补充。对于客户的基本信息，我们从用户的地理位置信息, 年龄性别会员等级等身份信息进行了相关分组操作和可视化处理; 对于用户的飞行记录, 我们主要从飞行时间, 飞行次数和基本积分等角度进行数据描述和可视化,

并分析其与身份信息如并研究不同组用户之间的层次差异。

在此基础上，我们建立了基于 LRFMC 的客户价值识别模型和客户流失模型。我们首先使用 KMeans 聚类方法对用户进行分类，并将其分为不同客户。然后使用常用的机器学习方法对于客户价值进行了预测，达到了很高的预测精度。而对于老客户的客户流失模型, 我们通过老客户的乘机次数进行分类判断，预测精度较高。而这些分类的识别与预测为更好地发掘用户的实际价值提供了参考和决策价值。因此，航空公司可以参考针对积分兑换, 会员等级相关的建议, 根据这些对于用户行为模式的分类制定相应的差异化策略，尽可能达到挽留住老用户, 发展新用户, 留住现有的发展保持用户等。

参考资料

[1] Haag S, Cummings M. Management information systems for the information age[M]. McGraw-Hill, Inc., 2009.

[2] 刘攀. 基于数据挖掘的航空公司客户价值建模 [D]. 华南理工大学. 2010.6

[3] 刘才雄. 基于多元统计方法的航空公司客户价值研究 [D]. 浙江工商大学. 2018.3

[4] 张良均等. Python 数据分析与挖掘实战 [M]. 机械工业出版社. 2016

[5] (美) 迈克尔. 波特. 竞争优势 [M]. 北京. 华夏出版社.2013

致谢与说明

本文中使用的数据是老师在网络学堂上提供的**国内某航空公司会员数据.xlsx**[4]。并部分参考[4]中关于数据挖掘模型的定义和处理建议。这里感谢毛老师一学期的辛勤授课和线上教学，并且希望老师以后的授课教学中最好能在课堂上加入对于较大规模基于编程软件的数据分析实践有一个稍微详细一点的纲领性指导 (预处理, 数据提取, 可视化, 数据挖掘等部分)。谢谢老师！

本数据集主要采用的分析工具为 python 3.7.4，主要使用的分析处理宏包有 pandas, numpy; 数据可视化宏包有 pyecharts, seaborn, matplotlib 等。所有程序均在一台 Intel(R) Core(TM) i7-7500U CPU @2.70GHz PC 机上运行。

[4]查询资料后发现, 原始数据来源于“泰迪杯”全国大学生数据挖掘竞赛网站 (http://www.tipdm.org/ts/661.jhtml), 对应航空公司为中国南方航空集团有限公司。

7 附录

其他文件说明：本文的程序代码使用 Jupyter Notebook 编写, 文件内容为 **project.ipynb**。关于本文详细的图可参考原始 img 文件夹下 echarts 生成的对应 html 文件。本文使用的国家代码数据集见 **country_code.csv** 等。

由于篇幅有限，这里仅列出关于 LRFMC 模型根据年龄段的细分模型。我们按照之前年龄等宽离散化的结果, 将 17-37 岁的会员记作青年人 (共 21098 人)，38-58 岁的会员记作壮年 (共 36429 人), 59-89 岁的会员记作老年人 (407 人)。随后我们同样运行 cluster_number = 5 的 KMeans 算法，分别针对少/壮/老三类人给出如下的图示结果:[5]。

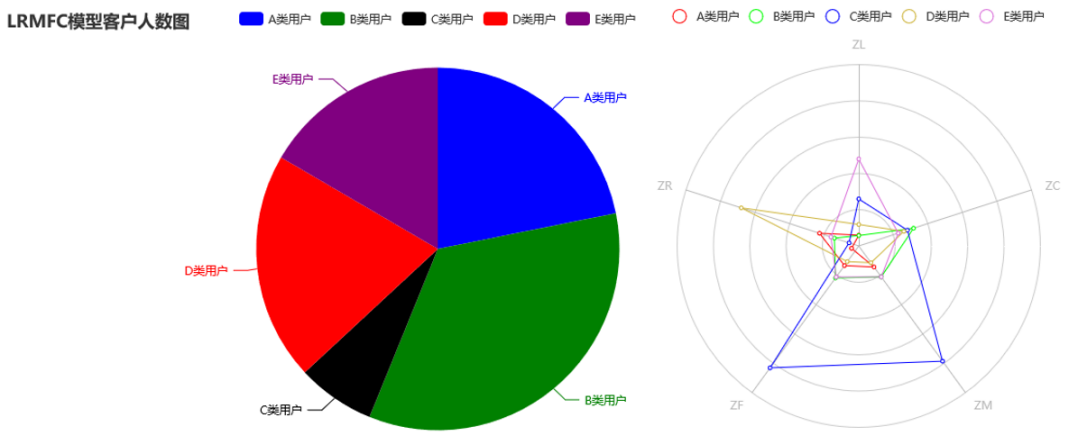

图 35: 航空公司 LRFMC 模型客户价值分类与特征图-青年

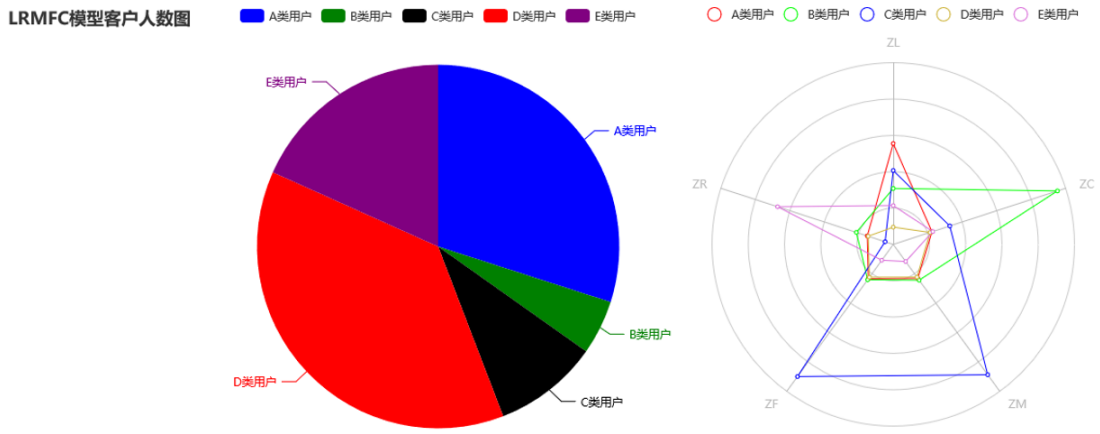

图 36: 航空公司 LRFMC 模型客户价值分类与特征图-壮年

[5]注意到由于 KMeans 算法随机生成初始点的特点, 本节所指的 A 类用户与之前正文的 A 类用户可能不是同一类用户

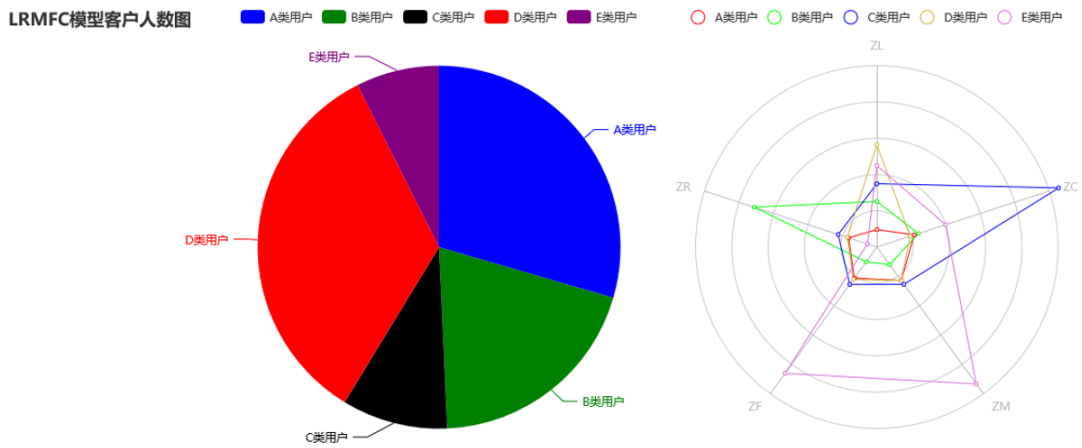

图 37: 航空公司 LRFMC 模型客户价值分类与特征图-老年

由上图可以看出, 相比年龄较大的会员群体, 青年中重要保持会员占比较低 (1-C,2-C,3-E)。但是其用户中一般与低价值会员相对较少，从聚类结果看五个聚类中心 ZC 较为平衡。而在壮年中他们的高价值用户和低价值用户占比较多。而对于性别的细分模型同样也能分析出类似的差异性结论，在这里不详细阐述。