分类号：
密　级：
UDC：

贵州财经大学

硕士学位论文

论文题目：航空公司客户分类及流失预测

专业名称：应用统计

研究方向：大数据分析

学生姓名：申玉伟

学　　号：20161208151003

导师姓名：张文专

导师职称：教授

定稿时间：

中国·贵阳

贵州财经大学学位论文原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。因本学位论文引起的法律结果完全由本人承担。

特此声明

学位论文作者签名：申玉伟

2019年6月10日

学位论文版权使用授权书

本学位论文成果归贵州财经大学所有。本作者完全了解学校有关保留、使用学位论文的规定，学校有权保留送交论文的复印件和磁盘，允许论文被查阅和借阅。本人授权学校可以将学位论文的全部或部门内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文，同意学校将本人的学位论文提交清华大学中国学术期刊（光盘版）电子杂志社全文出版和编入CNKI《中国知识资源总库》，传播学位论文的全部或部分内容。（保密的学位论文在解密后使用本授权书）

学位论文作者签名：申玉伟　　指导教师签字：

签字日期：2019年6月10日　　签字日期：2019年6月10日

摘要

目前，中国航空运输业已经初步形成了以中国国航、东方航空、南方航空等三大航空公司为主导，多家航空公司并存的竞争格局。面对激烈的市场竞争，航空公司面临着旅客流失，竞争力下降和航空资源未充分利用等经营危机。在此背景下，本文利用航空公司的客户数据对客户进行分类，并为航空公司提供较好的客户流失预测模型。具体工作概括如下：

1.数据预处理与描述性统计分析。对客户信息进行数据清洗、异常值处理，获得建模所需要的初始数据。另外，对航空公司客户的基本资料、飞行信息和积分信息进行描述性统计分析。

2.对航空公司客户进行分类。根据航空公司客户行为的 5 个指标（LRFMC）数据， 首先用 K-means 聚类算法对客户进行聚类分群，对每个客户群进行特征分析。然后使用 AHP 确定各指标权重，对各类别客户的价值进行量化分析，弥补航空公司客户分类方法的不足。

3.建立航空公司客户流失预测模型。首先对航空公司流失客户做出定义，对建立模型所用的指标进行说明，使用 logistic 回归模型，对流失客户的特征进行分析，建立航空公司客户流失的 logistic 回归模型，对航空公司客户是否流失作出预测。根据集成学习算法中的随机森林和梯度提升树模型，识别出流失客户，找出影响航空客户是否流失的比较重要的特征变量。最后对三个客户流失预测模型的性能进行对比分析，通过模型评估指标和 ROC 曲线两方面的对比，得出使用航空公司客户数据对客户进行流失预测的最优模型为梯度提升树模型。

本文旨在通过对航空公司客户进行分类，使航空公司针对不同价值客户采取不同营销策略，实现利润最大化。改善客户流失问题，使航空公司维护自身市场，给航空公司带来高利润。

关键词：客户分类；K-means；流失预测；logistic 回归；随机森林；梯度提升树

Abstract

At present, air transport industry has formed a competition pattern in which Air China, Eastern Airlines, Southern Airlines and other three major airlines are dominant and many airlines coexist. Faced with fierce market competition, airlines are facing such operational crises as loss of passengers, decline of competitiveness and underutilization of aviation resources. Under this background, this thesis classifies customers by using airline's customer data, and provides better customer churn prediction model for airlines. The specific work is summarized as follows:

1. Data preprocessing and descriptive statistical analysis. Customer information is cleaned and outliers are processed to obtain the initial data needed for modeling. In addition, descriptive statistical analysis is made on the basic information, flight information and integral information of airline customers.

2. Classify airline customers. Firstly, according to five indicators of airline customer behavior (LRFMC) data, K-means clustering algorithm is used to cluster customers and analyze the characteristics of each customer group. Then it uses the analytic hierarchy process to determine the weight of customer behavior indicators, and quantifies the value of each category of customers to make up for the shortcomings of the airline customer classification method.

3. Establishment of airline customer churn prediction model. Firstly, it defines the customer churn of airlines, explains the indicators used to establish the model, establishes the logistic regression model of customer churn, analyses the characteristics of the customer churn, and predicts whether the customer churn of airlines. According to the stochastic forest and gradient lifting tree model in ensemble learning algorithm, the lost customers are identified, and the more important characteristic variables affecting the loss of aviation customers are found. Finally, the performance of three customer churn prediction models is compared and analyzed. By comparing the model evaluation index and ROC curve, the optimal model of using airline customer data to predict customer churn is the gradient lifting tree model.

The purpose of this thesis is to classify airline customers so that airlines can adopt different marketing strategies for different value customers and maximize profits. Improve customer churn, so that airlines maintain their own market, and bring high profits to airlines.

Key words: customer classification; K-means; loss prediction; logistic regression; random forest; gradient lifting tree

目录

1 绪论

2011 年至 2018 年，中国全部航空公司营业收入总额由 3532 亿元增长至 8750 亿元。为了进一步提高航空公司的上座率，充分利用航空资源，使航空公司的效益得到提升。本文从客户关系的角度，通过研究航空公司客户分类和客户流失这两个方面来提高航空公司效益。

1.1 研究目的及意义

1.1.1 研究目的

航空公司在客户关系管理所出现的问题中，最为关键的是客户分类和客户流失问题。

客户分类逐渐成为亟待解决的关键问题之一，因为企业管理客户关系时，只有先对客户进行准确的分类，才能对营销资源做好优化分配。在日益激烈的市场竞争中，各航空公司纷纷推出了许多优惠的营销方式，以此来吸引更多的客户，竞争力下降，航空资源利用不足等已成为航空公司面临的主要运营危机。通过客户消费行为数据，对客户进行分群和特征分析，并对不同客户群的价值进行量化比较，航空公司据此制定相应的营销策略，为不同的客户群提供个性化的客户服务。

客户流失会对航空公司的利润增长产生非常不利的影响，其影响仅次于公司规模、市场份额和单位成本等[1]。流失一个客户比获得一个新客户的损失来的大，针对客户流失这个问题，本文建立航空公司客户流失模型，预测客户是否有流失倾向。

1.1.2 研究意义

使用数据挖掘方法，依据航空公司客户的消费行为特征，对航空公司的客户进行分类，识别出航空公司客户中的无价值客户和高价值客户。然后，根据客户的价值，航空公司对高价值和无价值的客户采用不同的销售策略，并制定出良好的、具有个性的服务计划。在营销资源有限的情况下，将目标瞄准高价值客户，从而实现航空公司利润的最大化。

改善航空公司客户流失的问题，使客户的满意度和忠诚度得到提高，航空公司在面对越来越激烈的竞争时能够维护好自己的市场。另外，通过客户流失预测，能够帮助航空公司做出持续的改进，给航空公司带来高利润。

1.2 研究现状

1.2.1 客户分类研究现状

基于行为属性进行客户分类：李琦，崔睿[2]依据客户的基本特征和交易行为，运用 RFM 模型构建客户分类的指标体系，使用 SOM 模型对网络保险业客户进行聚类分析，并对分类出的六类客户，选取出价值最高的两类客户，对其进行忠诚度分析并提出相应的营销策略建议。郎丽君，胡如夫[3]等为了更好地研究客户行为，对传统的 RFM 指标进行了扩充，在客户分类指标选取上采用基于 AHP 的 RFM 指标确定权重方法，通过自组织映射（SOM）和支持向量机（SVM）对选取出来的指标进行客户分类。王园[4]提出了一种基于聚类、规则提取和决策树方法的混合计算方法来预测以客户为中心的公司新客户的细分。首先，应用 K-均值算法对公司过去的客户进行基于购买行为的聚类。然后，提出了一种基于滤波和多属性决策的混合特征选择方法。最后，根据客户的特点，采用决策树分析法，挖掘出 IF 规则。所提出的方法被应用于证券业的案例研究，以便预测潜在盈利线索，并概述客户可用的最有影响力的特征，以便执行该预测。结果验证了所提出的方法的有效性和适用性，以处理现实生活中的情况。宋中山，周腾，周晶平[5]提出一种改进的自适应蚁群聚类算法，使用这种算法对客户进行分类，并与传统的 K-means 方法进行对比，结果表明这种算法对客户分类的结果更好。曾小青，徐秦 [6] 等用多指标代替传统的 RFM 指标，采用因子分析法分析多指标中的各个指标的权重，然后对客户进行有效的分类。颜书云[7] 提出了一种基于 GMDH 的半监督特征选择（GMDH-SSFS）算法，并将其应用于客户特征选择。此外，还考虑了特征和类标签之间的关系，以及特征选择过程中特征之间的关系。GMDH-SSFS 模型主要包括三个阶段：1）基于带类标签的数据集 L 训练 N 个基本分类模型；2）在无类标签的数据集 U 中选择性地标记样本，并将其添加到 L；3）基于新的训练集 L 训练 GMDH 神经网络，并选择 t 最佳特征子集 FS。根据选择出的特征，利用上海市某银行的信用卡数据,将客户分为高、中、低三个类别。张劲松，江波[8]选用机场旅客的行为数据为样本，使用决策树算法，对民航客户进行价值分类取得了良好的分类效果，然后根据分类结果分析了价值客户的主要特征。王光辉，张晓光、赵艳芹[9]用遗传算法优化 BP 神经网络的初始权值和阈值，有效地解释了非线性特征空间中的各特征，并确定用户类别。宋才华，蓝源娟，王永才，翟鸿荣，李滨涛[10]使用改进的熵权法，计算出客户综合价值指标的综合权重，提出了供电局客户分类的综合方法。胡晓雪，赵嵩正，吴楠[11]提出一种 SOM-DB-PAM 混合聚类算法，结合不同的分区聚类技术或分区和分层聚类技术，提出了一种高效的自底向上混合层次聚类（BHHC）技术，用来选择客户分类的原型。潘俊，王瑞琴[12]提

出了一种客户细分框架，这个框架是基于选择性聚类集成的，解决数据密集型企业的客户分类问题，首先将客户的特征划分为若干个子集，然后将客户对象聚类。高永梅，琚春华，邹江波[13]提出基于改进的多元逻辑回归对电信客户进行分类的方法。首先，使用 K-means 算法对大量的电信客户进行聚类，然后使用主成分分析法对客户特征属性进行降维，最后使用降维后的属性特征数据建立 logistic 回归模型对电信客户进行分类。邹轩[14]以电力客户缴费时间和缴费渠道为分类变量,将居民电力客户细分为欠费客户，拖费客户和正常缴费客户。李伟，秦鹏，胡广勤，张毓福[15]使用 R 和 Hadoop 集成编程环境（RHIPE）作为一个灵活的、可扩展的环境，RHIPE 支持对电网客户数据集的探索性数据分析，允许开发反映由实际数据表示的事件特征的数据清理和事件分类方法，而不是依赖于理论模型，选择 K-means 算法进行聚类，提高了分类的准确度。孙晓琳，姚波，陈瑜[16]对具有离群数据资产的数据进行客户分类问题,根据客户的交易频率，交易的产品或服务的种类，交易金额，客户年龄，采用先聚类再分类的方法，将离群客户分为 4 类。

依据价值属性的客户分类：陈明亮[17]将客户全生命周期利润（简称 CLP）作为判别客户对公司价值大小的标准，提出基于 CLP 的客户细分方法。陈明亮将客户价值分为当前价值和潜力价值两个维度，每个维度又可分为高低两档，最终将客户价值分为四类：当前价值和潜力价值均低、当前价值低但潜力价值高、当前价值高但潜力价值低、当前价值和潜力价值均高。朱明英，刑豫，王海霞，王保忠[18]通过研究在整个客户生命周期长度内企业从某一客户处获得的利润流的总现值,来计算客户的价值，使用 CLV 模型对电信客户数据进行分类。Achim Walter,Hans Georg Gemunden，Thomans Ritter[19]将客户价值划分为社会价值、直接价值和间接价值，并进行客户细分。李菊,王星,徐山[20]提出了基于粗糙集的客户分类预测模型，可以基于一些价值属性进行客户分类，降低了决策者的复杂度，该模型可以帮助企业提前预测新客户或潜在客户价值水平，最后的实证分析结果显示，该分类方法有效降低了数据计算的复杂度，提高了客户预测的准确性。目前，客户分类得到了广泛的应用，但针对航空公司这一市场的旅客细分研究还比较贫乏。

1.2.2 客户流失研究现状

AdnanAmin，Feras Al-Obeidat，Babar Shah[21] 等提出了一种新的 CCP 方法。根据距离因子将数据集分组成不同的区域，然后将其分为两类：高确定性数据、低确定性数据，用于预测呈现流失和非流失行为的客户。对不同的公开可用的电信工业（TCI）数据集使用不同的最新评估措施（例如，准确性、f-测量、

精度和召回率）表明：（1）距离因子与分类器的确定性密切相关（2）分类器在距离因子值较大的区域（即具有高确定性的客户搅动和非搅动）中获得比距离因子值较小的区域（即具有低确定性的客户搅动和非搅动）中更高的精度。Farid Shirazi,Mahbobeh Mohammadi[22]利用大数据量，包括结构化档案数据，结合网上网页、网站访问次数、电话会话日志等非结构化数据，构建一个预测性流失模型。它还考察了客户行为的不同方面对搅动决策的影响。应用 Hadoop 平台上的 Datameer 大数据分析工具和 SAS 商业智能系统的预测技术，建立了客户流失预测模型。王重仁，韩冬梅[23]结合成本敏感学习与极值梯度提升（XGBoost），提出一种成本敏感型增强型树状评估模型，以提高识别潜在流失客户的能力，提出一个将投资组合优化问题转化为整数线性规划的投资组合分配模型，作为非流失客户的决策支持系统，与以往的研究不同，作者的研究主要通过基于精度的度量来评估模型，并提出反映不同拒绝率下的 ARR 的新度量（ARR 曲线）。对用户社交数据进行了检验，数据分析表明，解释性变量在预测客户流失能力方面存在差异。实验结果表明，ARR 曲线下的面积不是评价模型的重要指标。Kuanchin Chen,Ya-Han Hu,Yi-Cheng Hsieh[24]探讨物流企业客户价值模型的长度、最近期、频率、货币与利润对客户流失的预测效果。与主流客户流失研究相比，这种独特的上下文具有有用的业务含义，其中单个客户（而不是业务客户）是主要的焦点。结果表明，五个变量对客户流失有不同的影响。具体而言，长度、最近期和货币变量对波动有显著影响，而频率变量只有在前三个变量的可变性受到限制时才成为最高预测因子。利润变量从来没有成为一个重要的预测因子。程昊，樊重俊[25]选取电商客户的年龄，性别，会龄、订单金额等指标，建立客户流失模型并对其进行评估。卢美琴，吴传威[26]结合商业银行的业务现状，用决策树算法对银行客户建立流失预警模型。卢光跃，王航龙，李创创[27]等提出公司客户流失预测(CCCP)是一个研究领域，可以使用公司的数据成功地预测客户流失。为了支持 CCCP，在建立 CCCP 模型之前，将公司数据转换为正态分布。然而，目前还不清楚哪种数据转换方法在 CCCP 中是最有效的。此外，在电信部门，数据转换方法对使用不同分类器的 CCCP 模型性能的影响还没有被全面地探讨。研究中，使用资料转换方法设计了一个 CCCP 的模型，并且提供广泛的比较来验证这些转换方法在 CCCP 中的影响，并且评估 K 近邻和支持向量机相结合的方法的性能。用于使用上述数据转换方法的电信部门客户流失预测。在公开的与电信部门有关的数据集上进行了实验，使用这种方法能够提高总体评价指标。总之，在电信业和金融行业，客户流失预测得到了广泛的应用。

1.3 本文主要研究内容及技术路线

1.3.1 主要研究内容

本文主要依据航空公司客户的相关数据，进行两方面研究。第一：选取航空公司客户的消费行为指标，对航空公司客户分类，对每类客户进行消费行为特征分析，并使用 AHP 对各类客户的价值进行量化分析，直观的比较各类客户的价值。第二：根据 logistic 回归模型，分析流失客户特征。构建客户流失预测的 logistic 回归模型、随机森林模型和梯度提升树模型，对航空公司客户进行流失预测，最后对模型进行对比分析评价，选取出针对航空公司客户数据最优的客户流失预测模型。本文的主要工作概述如下：

第一章绪论。主要包括研究背景和意义，关于客户分类和客户流失预测的研究现状、本文研究的技术路线和创新之处。

第二章航空公司客户数据预处理和描述性统计分析。对客户信息进行数据清洗、异常值处理，获得建模所需要的初始数据。另外，对航空公司客户的基本资料信息、飞行信息和积分信息数据进行描述性统计分析。

第三章航空公司客户分类。根据航空公司客户关于客户行为的 5 个指标（LRFMC）数据，首先用 K-means 聚类算法对客户进行聚类分群，对每个客户群进行特征分析。然后使用 AHP 确定客户行为指标权重，对各类别客户价值进行量化分析，弥补航空公司客户分类方法的不足。

第四构建预测航空公司客户流失的模型。根据 logistic 模型，分析流失客户特征。使用统计学中的经典分类模型 logistic 回归模型，集成学习算法中的随机森林和梯度提升树模型，来预测出航空公司的流失客户。最后对三个流失客户模型的预测结果进行对比分析，通过模型评估指标和模型 ROC 曲线两方面的对比，得出最优的客户流失预测模型为航空公司客户流失的梯度提升树模型。

第五章结论与展望。对本文所做的工作和得出的结论进行总结，然后对后续的航空公司客户研究方法、内容给予参考建议。

1.3.2 技术路线

本文选取航空公司客户消费行为指标，用 K-means 算法对航空公司客户进行聚类，得到五类客户群并分析每个客户群的客户特征，再使用层次分析法确定每个指标的权重，对每类客户群的价值进行量化分析，得出价值得分并进行排序。建立航空公司客户流失的 logistic 回归模型、随机森林模型和梯度提升树模型，对航空公司客户是否有“流失倾向”进行客户分类，对三个模型的分类效果进行对比分析，得出能够较好的预测航空公司流失客户的模型。下图为本文的技术路

线展示：

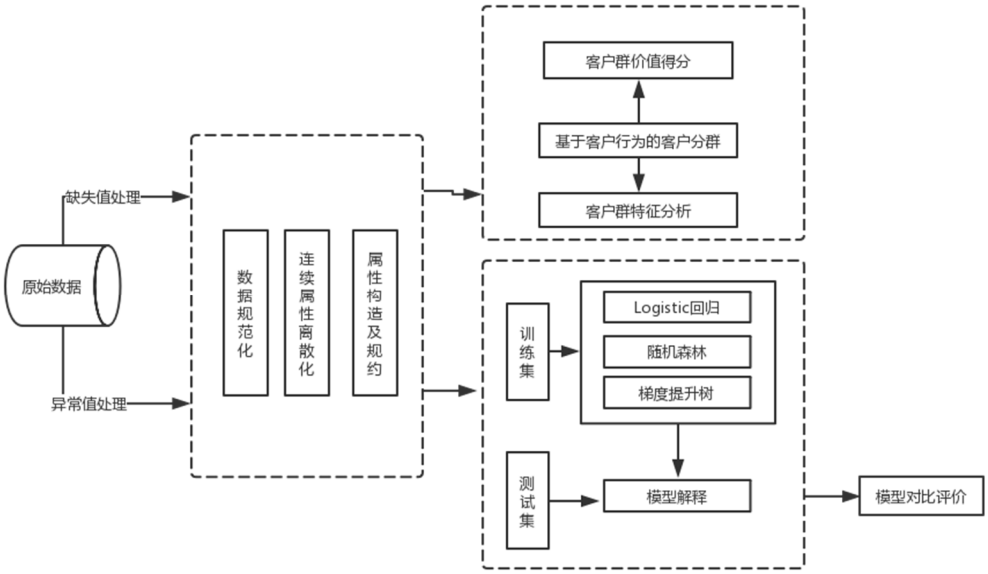


图 1.1 本文技术路线

1.3.3 创新之处

(1)作为重要的数据挖掘技术，聚类被广泛的应用在市场细分问题中。本文使用聚类方法将航空公司客户进行分类，并使用层次分析法确定客户行为指标的权重，对各类别客户价值进行量化分析，弥补航空公司客户分类方法的不足，帮助航空公司制定更可行的客户策略。

(2)对于航空公司客户信息数据，鲜有人对客户进行流失预测。本文分别建立航空公司客户流失的 logistic 回归模型、随机森林模型和梯度提升树模型，识别流失客户，分析流失客户特征。从而改善航空公司客户流失的问题，给航空公司带来高利润。

2 数据预处理及描述性统计分析

2.1 数据说明

本文使用国内某航空公司的62988名客户数据作为原始数据集，每条数据包括 63 个属性，包括客户基本资料、用户飞行信息和积分信息数据(2 年内)。本文所使用的工具主要有 Python、R、Excel。

1.客户基本资料

客户基本资料包括航空公司客户会员卡号、性别、年龄、工作地所在城市、工作地所在省份、工作地所在国家、入会时间和会员卡级别。

2. 用户飞行信息

用户飞行信息包括飞行次数、观测窗口总飞行公里数、观测窗口最后一次飞行日期、观测窗口季度平均飞行次数、平均乘机时间间隔、观测窗口内最大乘机间隔、平均折扣率、第一年总票价、第二年总票价等。

3. 用户积分信息

用户积分信息包括第 1 年里程积分、第 2 年里程积分、观测窗口总基本积分、观测窗口季度平均积分、观测窗口中总精英积分、观测窗口中其他积分、积分兑换次数等。

2.2 数据预处理

一般来说，原始数据集具有不完整、有噪声、不一致的特点，而数据清洗就是针对这些特点，将数据集补充完整，光滑噪声，把不一致纠正过来。下面对航空公司的会员数据进行数据清洗。

表 2.1 航空公司客户数据探索分析结果表

属性名称	空值数	最大值	最小值
EXPENSE_SUM_YR_1	551	239560	0
EXPENSE_SUM_YR_2	138	234188	0
FLIGHT_COUNT	0	213	2
FLIGHT_COUNT_QTR_1	0	39	0
FLIGHT_COUNT_QTR_2	0	31	0
FLIGHT_COUNT_QTR_3	0	32	0
…	…	…	…
EXCHANGE_COUNT	0	46	6
avg_discount	0	1.5	0

经分析，有 956 条数据存在缺失值，另外数据中存在票价最小值为 0，但折扣率和飞行公里数却不为 0 的异常数据。数据集中有 1140 个样本存在异常值和缺失值，只占原始数据集的 1.8%，对这些样本直接做删除处理，处理后数据有 61848 个样本。

2.3 描述性统计分析

下面对航空公司客户的基本资料信息、飞行信息和积分信息数据进行描述性统计分析，用统计图初步展示数据。

1.基本资料信息

图 2.1 会员人数

由图 2.1 可知，航空公司的客户中，男性会员有 47297 人，占所有会员人数的 76.5%，女性会员有 14551 人，占所有会员人数的 23.5%。

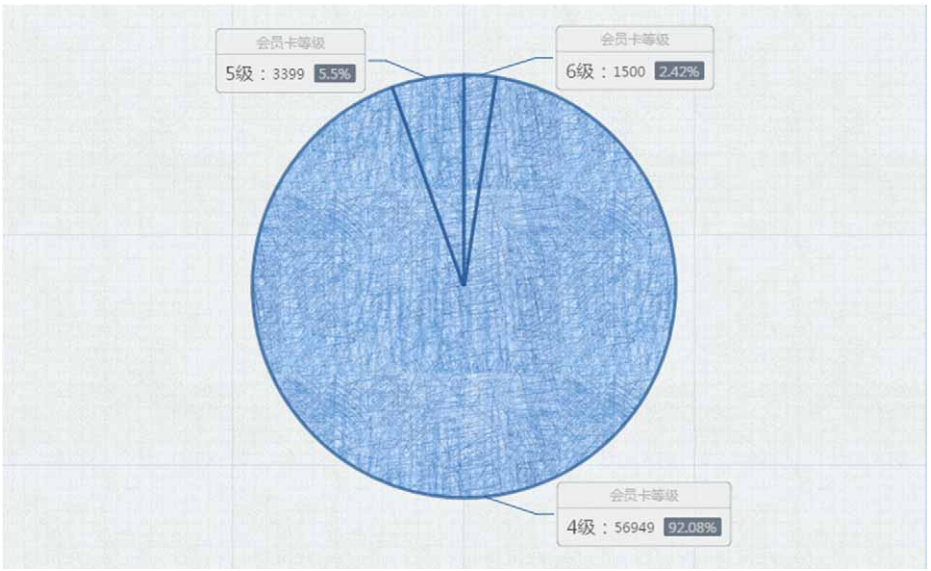


图 2.2 会员等级

由图 2.2 可知，会员等级为 4 级的人数最多，有 56949 人，占全部会员人数的 92.08%，会员卡等级为 5 级的人数为 3399 人，占全部会员人数的 5.5%，会员卡等级为 6 级的人数有 1500 人，占全部会员人数的 2.42%。

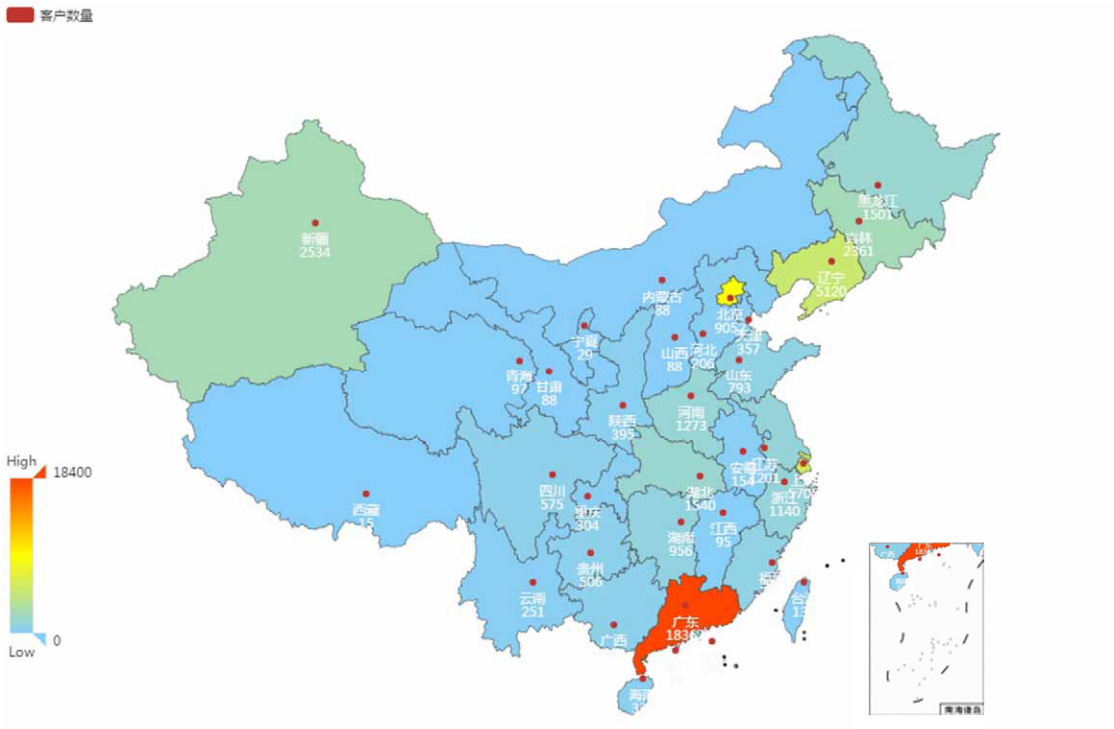


图 2.3 客户工作所在地分布图

航空公司的客户中，工作所在地在国内（含港澳台）的人数有 57669 人，工作所在地在国外的人数有 4179 人。工作地在广东的客户人数最多，高达 18367 人，其次是北京，有 9052 人，工作地在西藏的客户人数最少，只有 15 人。

2.飞行信息

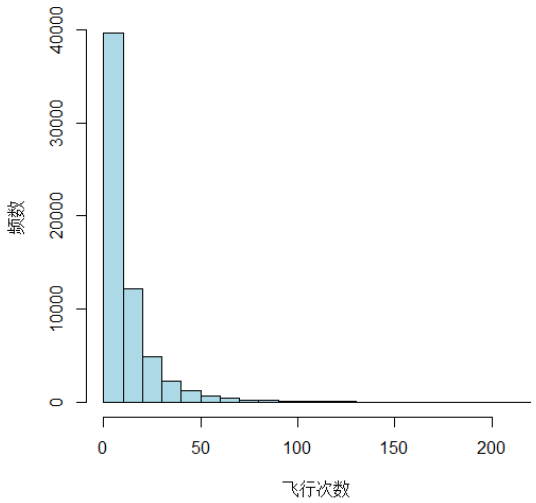


图 2.4 飞行次数频数分布直方图

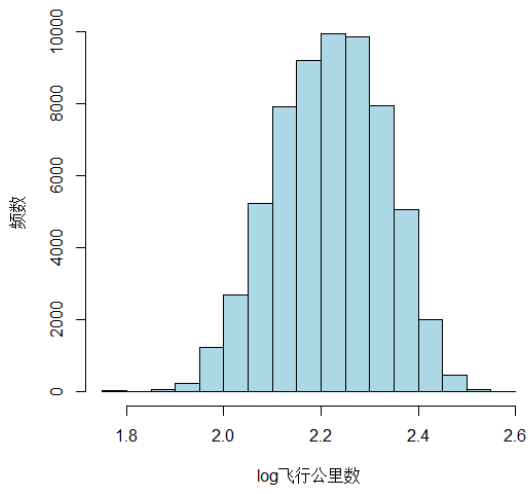
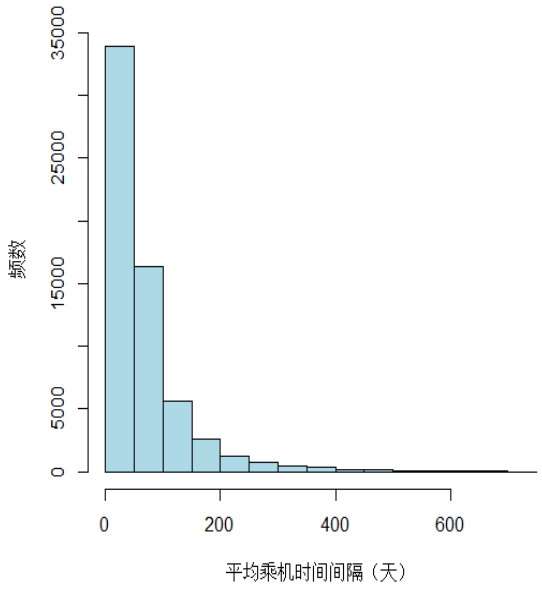


图 2.5 log 飞行公里数频数分布直方图

由图 2.4 可知，观测窗口（2 年）内，将近 4 万航空公司的客户飞行次数在 10 次内。由图 2.5 可知，Log 飞行公里数基本呈现对称分布，最小飞行公里数为 368 公里，最大值为 235687 公里。



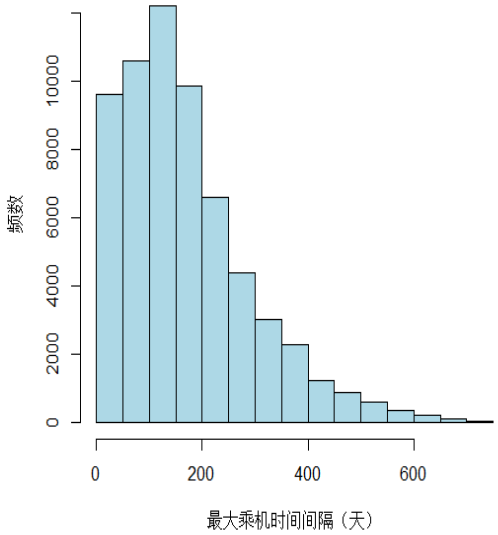
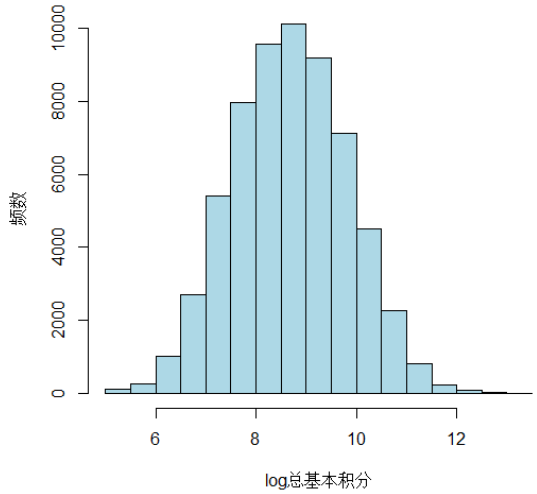


图 2.6 平均时间间隔频数分布直方图　　图 2.7 最大时间间隔频数分布直方图

由图 2.6 可知，平均乘机时间间隔在 50 天内的乘客人数最多，有 33893 人。由图 2.7 可知最大乘机时间间隔集中在 100 天左右，最大乘机时间间隔在 200 天内的乘客有 42271 人。

3.积分信息



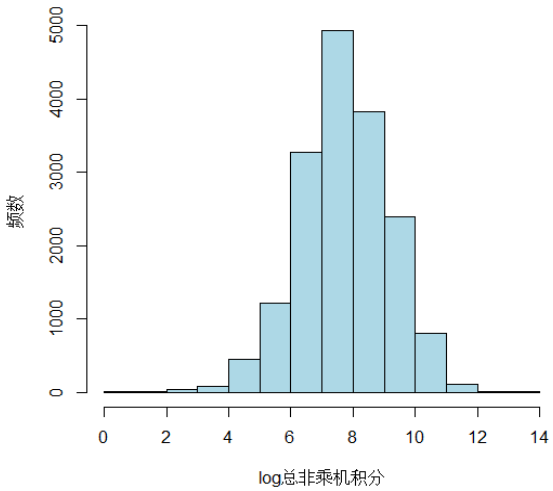
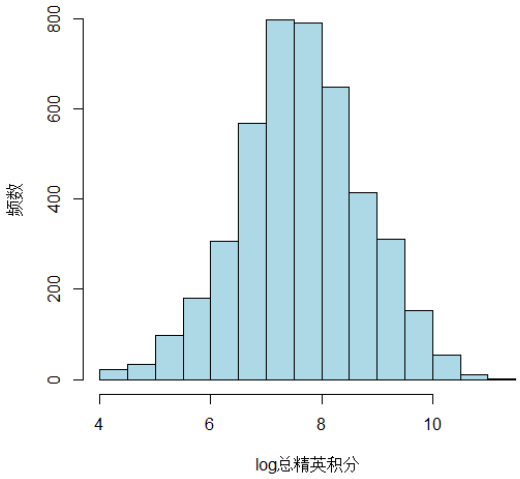


图 2.8 log 总基本积分频数分布直方图　图 2.9 log 总非乘机积分频数分布直方图

由图 2.8 可知，log 总基本积分基本呈现对称分布，最小总基本积分为 30，最大总基本积分为 286164。由图 2.9 可知，最小总非乘机积分为 0，最大总非乘机积分为 984938。



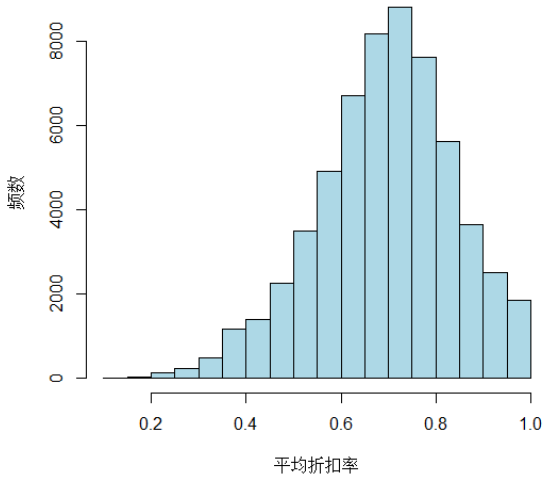


图 2.10 log 总精英积分频数分布直方图　　图 2.11 平均折扣率频数分布直方图

由图 2.10 可知，log 总精英积分基本呈现对称分布，最小总精英积分为 0，最大总精英积分为 74460。由图 2.11 可知，平均折扣率集中在区间[0.5,0.9]，平均折扣率在区间[0.7,0.75]的人数最多，有 9909 人。

3 航空公司客户分类及特征分析

本章研究航空公司客户关系管理中一个关键问题——客户分类问题，即将航空公司的客户进行聚类分群。本章根据航空公司客户的消费行为指标（LRFMC）数据， 首先用 K-means 聚类算法对客户进行聚类分群，对每个客户群进行特征分析。然后采用层次分析法确定每个特征的权重，得出每个客户群价值的得分，对客户群进行排名。

3.1 分类算法的选择

航空公司客户分类问题，即将客户进行聚类分群。聚类，是无监督问题的一个代表，在没有标签数据的情况下将数据分成不同的集合，通俗的讲就是把相似的归为一组，簇是数据对象的集合。常用的聚类算法有：

1. 划分聚类：随机选取k个聚类中心，计算各组样本到中心的距离，不断改变聚类中心，再次计算距离，一直迭代下去减小距离，直到聚类中心不再改变。

2. 层次聚类：起初将每个样本点都当做簇，按照接近度，簇与簇进行组合。

3. 基于密度聚类：密度聚类算法从样本密度的角度来考察样本之间的可连接性，并基于可连接样本不断拓展聚类簇,以获得最终的聚类结果[28]。

4. 基于网格聚类：把对象量化成有限个单元，形成网格结构，在网格结构上进行聚类。

王丽菊[29]也对上述四种聚类算法进行了介绍，因 K-means 聚类算法使用广泛且算法简单，处理大量的数据时速度快，使用 K-means 聚类算法对航空旅客进行划分，是适合做航空客户数据的聚类的算法。K-means 聚类算法使用距离作为相似性的评估指标，当两个对象之间的距离越近时，对象间的相似度也就越大。

3.2 K-means 聚类算法介绍

3.2.1 K-means 聚类算法过程

输入：

K:簇的数目

D：包含n个对象的数据集。

输出：K个簇的集合。

方法：

（1）从数据集中随机地选择k个对象，以这k个点作为聚类中心；

（2）计算出数据集中每个样本到k个聚类中心的距离，选择最近的聚类中心，将该对象归入这个类中；

（3）在分配了数据集中的所有对象之后，再重新计算k个簇的聚类中心；

（4）如果这次的聚类中心与前一次的k个聚类中心有发生变化，则重复步骤（2），否则继续下一步；

（5）当k个聚类中心不再发生改变时，停止聚类，输出结果。

3.2.2 数据类型与相似性的度量

（1）连续属性

对于连续属性，如果每个变量的取值范围较大，就要对每个属性进行零–均值规范，这样做可以消除不同数量级数据对聚类结果带来的影响。假设$d(e_i,x)$表示样本和簇之间的距离，$d(e_i,x)$越小越好；$d(e_i,e_j)$表示簇与簇之间的距离，$d(e_i,e_j)$越大越好。

n个样本的数据矩阵如下：

$$\begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} \tag{3.1}$$

欧几里得距离的计算公式为：

$$d(i,j)=\sqrt{(x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2+...+(x_{ip}-x_{jp})^2} \tag{3.2}$$

曼哈顿距离的计算公式为：

$$d(i,j)=\left|x_{i1}-x_{j1}\right|+\left|x_{i2}-x_{j2}\right|+...+\left|x_{ip}-x_{jp}\right| \tag{3.3}$$

闵可夫斯基距离的计算公式为：

$$d(i,j)=\sqrt[q]{(\left|x_{i1}-x_{j1}\right|)^q+(\left|x_{i2}-x_{j2}\right|)^q+...+(\left|x_{ip}-x_{jp}\right|)^q} \tag{3.4}$$

（2）文档数据

当聚类目标是文档数据时，先将其整理成文档–词矩阵格式，如表所示。

表 3.1 文档-词矩阵

	lost	*win*	*team*	*score*	*music*	*sad*		*coach*
文档一	14	2	8	0	8	10	…	6
文档二	1	13	3	4	1	4	…	7
文档三	9	6	7	7	3	8	…	5

两个文档之间的相似度的计算公式为：

$$d(i,j)=\cos(i,j)=\frac{\vec{i}g\vec{j}}{\left|\vec{i}\right|g\left|\vec{j}\right|} \tag{3.5}$$

3.2.3 目标函数

选择误差平方和（*SSE*）较小的分类结果。

连续属性的*SSE*计算公式为：

$$SSE = \sum_{i=1}^{K} \sum_{x \in E_i} dist(e_i, x)^2 \tag{3.6}$$

文档数据的*SSE*计算公式为：

$$SSE = \sum_{i=1}^{K} \sum_{x \in E_i} \cos ine(e_i, x)^2 \tag{3.7}$$

簇E_i的聚类中心e_i计算公式为：

$$e_i = \frac{1}{n_i} \sum_{x \in E_i} x \tag{3.8}$$

符号含义如表所示：

表 3.2 符号表

符号	含义
K	聚类簇的个数
x	对象（样本）
E_i	第i个簇
e_i	簇E_i的聚类中心
n	数据集中样本的个数
n_i	第i个簇中样本的个数

3.2.4 K-means 聚类算法的优势和劣势

K-means 算法既简单又很实用，其优、劣势总结见下表 3.3。

表 3.3 K-means 算法优劣、势总结

优势	劣势
原理简单，收敛速度快	k的选取不好把握
聚类效果较优	对于不是凸的数据集难以收敛
模型的可解释性较强	对于噪音和异常点比较敏感
调参只需要簇数k	结果最初k个中心影响

3.3 客户类型聚类过程及特征分析

3.3.1 指标选取

本文目标之一是客户分类，对客户价值进行识别，原始数据集中包括客户基

本信息、乘机信息和积分信息，特征属性太多，而且各属性不具有降维的特征。识别客户价值的模型中,应用最广泛的是RFM模型了。RFM模型使用了三个指标：客户最近消费的时间间隔（R），客户消费的频率（F），客户消费的金额（M）。R指标数值越小、F指标数值越大、M指标数值越大，客户价值越高。但在航空公司的业务中，即便两位航空客户的消费金额相同，一位客户乘坐的距离较远，但舱位等级低，而另一位客户乘坐距离近，但舱位等级高，客户对航空公司的价值是不等的。所以选取客户在观测窗口的累积飞行里程M和客户得折扣系数的平均值C来代替传统模型中的M。另外，考虑客户的入会时长在一定程度上也会影响客户价值，所以在模型中加入变量客户的关系长度L[30]。

3.3.2 属性规约与数据变换

选择与LRFMC指标相关的六个属性，并将其余属性删除。与选用指标相关的属性见下表3.4，数据属性构造见表3.5。

表3.4 与LRFMC指标相关的六个属性

序号	属性名称	属性解释
1	LOAD_TIME	观测窗口结束时间
2	AVG_DISCOUNT	平均折扣率
3	SEG_KM_SUM	观测窗口飞行里程
4	FLIGHT_COUNY	观测窗口飞行次数
5	DAYS FROM LAST_TO_END	最后一次乘机时间至观测窗口结束时长
6	FFP_DATE	入会时间

表3.5 指标计算方法

指标名称(单位)	计算方法
R（月）	DAYS FROM LAST_TO_END
F（次）	FLIGHT_COUNY
M（公里）	SEG_KM_SUM
C（无）	SEG_KM_SUM
L（月）	LOAD_TIME - FFP_DATE

下面对五个指标的数据进行分析，每个指标的最大值和最小值见表3.6，表中指标LRFMC的取值范围较大，需要对数据进行标准化处理。

表3.6 指标取值范围

指标名称	R	F	M	C	L
最小值	0.03	2	368	0.11	12.23
最大值	24.23	136	234721	1.5	114.63

表 3.7 标准化后指标取值范围

指标名称	ZR	ZF	ZM	ZC	ZL
最小值	-0.948	0.706	-0.804	-3.299	-1.324
最大值	3.079	14.267	26.799	4.208	2.3

3.3.3 模型构建及客户特征分析

采用 K-means 算法对航空公司的客户进行聚类分析，而参数 k 的取值是一个难点，假设 k 确定，就可以计算出不同 k 值时的误差平方和（ *SSE* ），其值越小越好。利用 *SSE* 找到合适的 k 值。

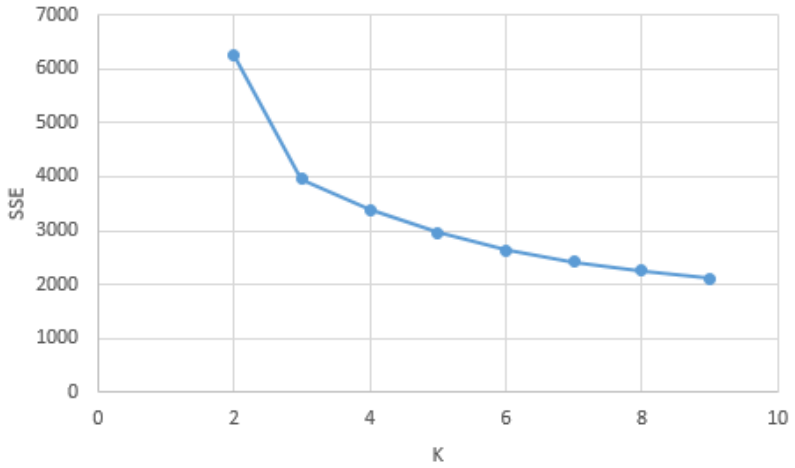


图 3.1 不同 k 值与 SSE 折线图

由图 3.1 可以看出，当 k 不断增大时， *SSE* 的值逐渐减小，并没有转折点出现。由上图可知，当 k 大于等于 5 时， k 值与 *SSE* 折线图中折线下降的趋势逐渐变得缓慢，为了达到好的分类效果且考虑效率的问题，这里选取 k=5，然后得出各类客户聚类中心的雷达图，通过雷达图的来分析选取的 k 值是否合适。

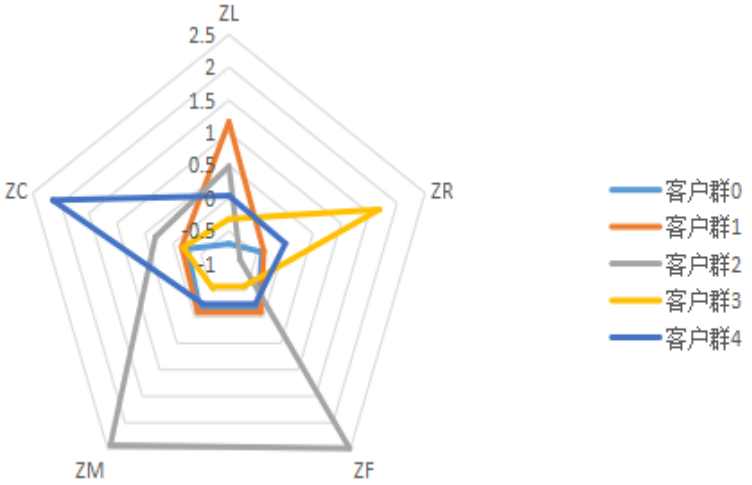


图 3.2 k=5 时客户群特征分析图

通过观察可知：当k值取为5时，分析的结果比较合理，而且分出的五类客户群都有自己的特点，而且各类客户特征不会相互重复。所以，$k=5$时，得到了较好的聚类效果。然后利用*Python*中的*Scikit-Learn*库下的聚类子库*sklearn.cluster*（采用欧氏距离）对航空公司会员数据构建K-means聚类（$k=5$），聚类结果如表3.8所示。

表3.8 客户聚类结果

聚类类别	聚类个数	聚类中心				
		ZL	ZR	ZF	ZM	ZC
客户群0	24589	-0.700	-0.415	-0.162	-0.161	-0.260
客户群1	15568	1.168	-0.377	-0.084	-0.092	-0.155
客户群2	12065	-0.316	1.684	-0.573	-0.536	-0.180
客户群3	4308	0.038	-0.003	-0.237	-0.243	2.147
客户群4	5318	0.488	-0.801	2.485	2.425	0.315

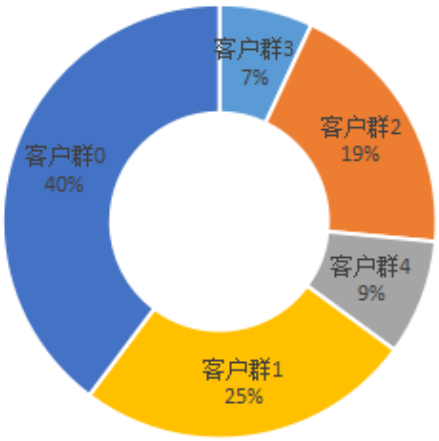


图3.3 各类客户所占比重

由图3.3可知，客户群0所占比重最大，占比40%。客户群0中的客户在观测窗口内的乘坐本公司航班的次数少，他们所乘航班的平均折扣率很低，且较长时间没有乘坐过本公司航班，这类客户是一般客户。

客户群1中的客户已经有较长时间没有乘坐过本公司航班，但是他们以往所乘坐本公司航班的平均折扣率、累计里程或乘机次数较高，客户的价值量变动方向具有很大的不确定性，这类客户是航空公司的重要挽留客户。

客户群2中的客户在观测窗口内乘坐本公司航班的次数低，累计里程数低，平均折扣率也低，而且已经有很长时间没有乘坐过本公司的航班，入会时长短，他们是低价值客户。

客户群3中的客户平均折扣率较高，近期有乘坐过本公司的航班，但入会时长短，乘机次数、里程均较低。这类客户的当前价值不高，但是发展潜力很大，他们是重要发展客户。

客户群 4 中的客户最近乘坐过本公司航班，观测窗口内乘坐次本公司航班次数多或者乘坐里程较高，平均折扣率较高，这类客户的价值很高，他们是重要保持客户。

根据各类客户特征，对客户群进行客户特征描述如图 3.4。

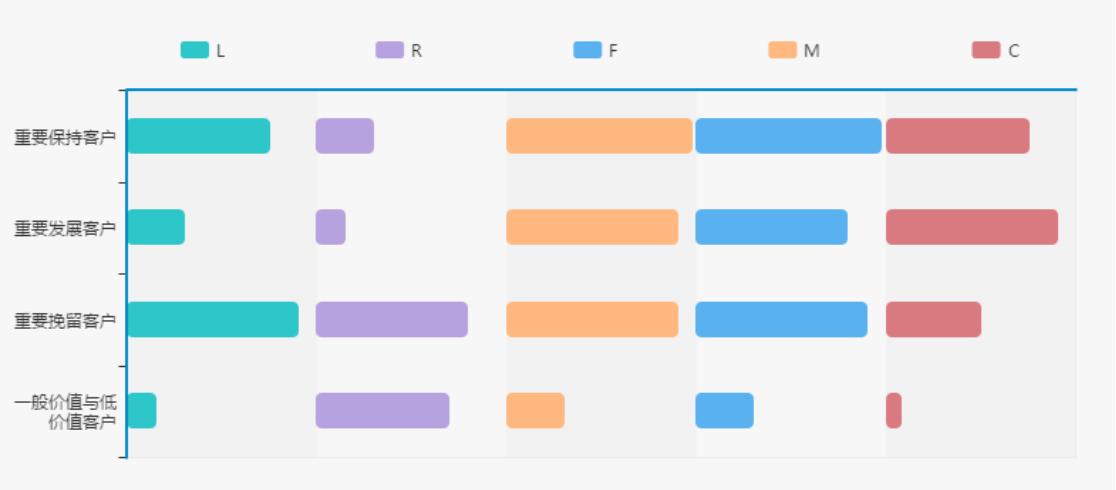


图 3.4 客户群特征描述图

3.4 客户价值排名

到底哪个客户群的价值高呢？为了对各类客户之间的价值有一个量化的比较，利用层次分析法得出每个特征的权重，计算每个客户群的价值得分，对客户价值进行排名。

首先需要对 5 个指标数据进行最小—最大规范化，转换公式为：

$$x^* = \frac{x - \min}{\max - \min} \tag{3.9}$$

由于 R 与顾客价值之间存在负相关的关系，所以将转化公式调整为：

$$x^* = \frac{\max - x}{\max - \min} \tag{3.10}$$

规范化后，上一小节中五个客户群的聚类中心的取值均在[0,1]区间内。

对于不同行业甚至是不同公司，模型中各指标的权重都存在着差异，所以分析指标权重时需要采用科学的方法。对此，以层次分析法确定指标的权重，首先得到下表的评价矩阵。

表 3.9 评价矩阵

	L	R	F	M	C
L	1	1/5	1	1/6	1/8
R	5	1	5	1	2
F	1	1/5	1	1/6	1/8
M	6	1	6	1	1
C	8	1/2	8	1	1

使用层次分析法得到 LRFMC 各个指标的权重 w_L, w_R, w_F, w_M, w_C，其中

$C_L^j C_R^j C_F^j C_M^j C_C^j$ 分别表示第 j 类客户聚类中心的 LRFMC 各个指标的值。C^j 是第 j 类客户的 LRFMC 各项指标加权后的总得分，计算公式为:

$$C^j = W_L C_L^j + W_R C_R^j + W_F C_F^j + W_M C_M^j + W_C C_C^j \quad (3.11)$$

根据层次分析法得到各个指标的权重为 $w_L = 0.038$，$w_R = 0.142$，$w_F = 0.364$，$w_M = 0.380$，$w_C = 0.077$。表 3.10 为各类客户价值得分的排序结果，从表中结果可知重要保持客户和重要挽留客户的客户价值较高。

表 3.10 五类客户的价值得分排名

客户群	客户价值得分	排名
一般客户	0.093	4
重要挽留客户	0.286	2
低价值客户	0.080	5
重要发展客户	0.273	3
重要保持客户	0.297	1

在对航空公司客户进行分类以后，航空公司对各个类别客户价值的差别进行进行量化分析，对客户的类别进行价值排序。这样不仅可以弥补航空公司客户分类方法的不足，还有助于航空公司制定更可行的客户政策。由于航空公司受到成本的制约，不可能采取无差别的个性化服务，只能将资源集中在少数几类对公司重要的客户上。按照总得分的排列情况，航空公司应优先把资源投放到总得分较高的客户身上。

4 航空公司客户流失预测模型

本章开始研究航空公司客户关系管理中的另一个关键问题——客户流失问题，即从航空公司客户原始数据中选择指标，建立航空公司客户流失预测模型。原始数据集中并未给出客户是否流失，所以首先要对客户是否流失作出定义，本文对流失客户作出如下定义：最后一次乘机时间至观测窗口末端时长大于等于观测窗口内最大乘机间隔。其中流失客户个数为 23558，非流失客户个数为 38290。

4.1 航空公司客户流失的 logistic 回归模型

Logistic 回归模型是经典的二分类模型，所以本小节选用航空公司客户流失 logistic 回归模型对航空公司客户流失进行预测。本节首先对 logistic 回归模型进行介绍，然后对建立模型所用的指标进行说明，并将全部客户数据进行 logistic 回归，根据回归系数分析流失客户特征。使用训练集数据建立航空公司客户流失的 logistic 回归模型，最后使用测试集对航空公司客户是否流失作出预测并对结果进行分析。

4.1.1 logistic 回归介绍

逻辑回归假定了目标 y 属于集合 $\{0,1\}$，是适用于分类的算法[31]。两分类问题，将正类记为 1，负类记为 0。在逻辑回归中，一个关键的假设是样本 x 属于正类的概率可以用下面的式子来表示：

$$p(y=1|x)=sig(w^T x) \tag{4.1}$$

这里 $sig()$ 称为 *sigmoid* 函数，其定义如下：

$$sig(t)=\frac{1}{1+\exp(-t)} \tag{4.2}$$

sigmoid 函数称为连接函数，$w^T x$ 是线性模型，通过连接函数来转换 $w^T x$ 的模型称为广义线性模型（*GLM*）。

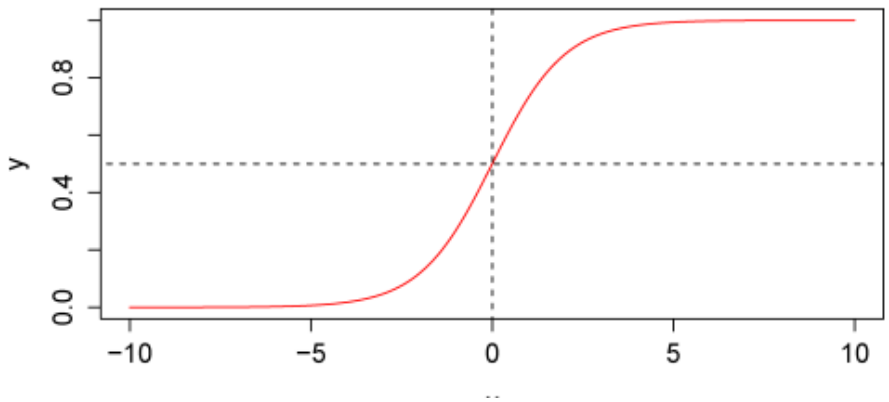


图 4.1 Sigmoid 函数的图像

图 4.1 给出了 *sigmoid* 函数的图像。*sigmoid* 函数的形状类似于 *S* 形。该函数的输入值取值范围为 $(-\infty,+\infty)$，但是其输出区间是 $(0,1)$。这样对于任意值的 $w^T x$，通过 *sigmoid* 函数，都能够得到 0 到 1 之间的概率。

对于参数的求法，可以使用极大似然估计来估计出各个参数。在逻辑回归中唯一的参数是 w。假设训练集为 $\{(x_1,y_1),(x_2,y_2)\cdots(x_n,y_n)\}$，引入极大似然函数，并找出最优解。依据对数似然函数，定义相应的损失函数。

首先考虑样本 (x_i,y_i)，使用 t_i 来标记样本 x_i 是正类的概率：

$$t_i = p(y_i = 1 \mid x_i) = sig(w^T x_i) \tag{4.3}$$

于是，给定 x_i 时，y_i 是 0 的概率为 $p(y_i = 0 \mid x_i) = p(y_i = 1 \mid x_i) = 1 - t_i$。可以将 y_i 的似然函数写为：

$$p(y_i \mid w) = t_i^{y_i}(1-t_i)^{(1-y_i)} \tag{4.4}$$

对于整个训练集 $\{(x_1,y_1),(x_2,y_2)\cdots(x_n,y_n)\}$ 而言，似然函数可以写为：

$$p(y_i \mid w) \prod_{i=1}^{n} t_i^{y_i}(1-t_i)^{(1-y_i)} \tag{4.5}$$

这里 $y = (y_1, y_2, \cdots, y_n)^T \in \square^n$。对于给定的训练集，可以最大化似然 $p(y \mid w)$，从而得到最优的 w。在逻辑回归中，引入损失函数——交叉熵（$cross-entropy$）损失函数：

$$E(w) = -\ln p(y \mid w) = -\sum_{i=1}^{n}(y_i \ln t_i + (1-y_i)\ln(1-t_i)) \tag{4.6}$$

先对似然函数取对数，然后再取其相反数，这样便得到交叉熵损失函数。因此，最大化似然函数 $\Leftrightarrow$ 最小化交叉熵损失函数。

在逻辑回归中，引入正则化项，用来控制模型的复杂度。

引入 $L1$ 范数：

$$E_1(w) = -\sum_{i=1}^{n}(y_i lnt_i + (1-y_i)ln(1-t_i)) + \lambda_1 \|w\|_1 \tag{4.7}$$

$\lambda_1 \geq 0$，是控制范数 $\|w\|_1$ 权重的系数。因为 $\|w\|_1$ 在 $w=0$ 点连续但不可导，所以在求解上更加复杂。

也可以引入 $L2$ 范数：

$$E_2(w) = -\sum_{i=1}^{n}(y_i lnt_i + (1-y_i)ln(1-t_i)) + \lambda_2 \|w\|_2^2 \tag{4.8}$$

$\lambda_2 \geq 0$，是控制范数权重的系数。前半部使用交叉熵损失函数来度量模型 f 在训练数据上的表现，后半部使用 $L2$ 范数来控制模型的复杂度。

也可以同时引入 $L1$ 范数和 $L2$ 范数：

$$E_{12}(w) = -\sum_{i=1}^{n}(y_i lnt_i + (1-y_i)ln(1-t_i)) + \lambda_1 \|w\|_1 + \lambda_2 \|w\|_2^2 \tag{4.9}$$

4.1.2 变量的选取与说明

本文数据集是来自国内某航空公司的会员数据，在进行过数据预处理后共有61848个样本（原始数据集样本数量为62988个）。航空公司客户流失的logistic回归模型中选择航空公司客户是否流失作为因变量，流失客户为1，非流失客户为0。自变量选择客户的基本资料信息中的5个指标、用户飞行信息中的14个指标和用户积分信息中的11个指标，详细的指标介绍见表4.1。另外本章中的其它两个流失预测模型——航空公司客户流失的随机森林和梯度提升树模型选取的变量与航空公司客户流失的logistic回归模型相同，下文中便不做赘述。

表4.1 指标说明表

变量类型		变量名	详细说明	取值范围
因变量		是否流失	定性变量	是(1)、否（0）
自变量	客户基本资料信息	入会时长	连续变量	12.23~114.63
		性别	定性变量	男、女
		会员卡级别	定性变量	4级、5级、6级
		工作国家	定性变量	国内、国外
		年龄	连续变量	6~92
自变量	用户飞行数据	2年内飞行次数	连续变量	2~213
		第1年票价总收入	连续变量	0~181233
		第2年票价总收入	连续变量	0~125500
		总飞行公里数	连续变量据	236~235687
		观测窗口总加权飞行公里数	连续变量	266~240843.6
		观测窗口季度平均飞行次数	连续变量	0.25~62.625
		观测窗口季度平均基本积分累积	连续变量	0~355770.5
		观察窗口内第一次乘机时间至入会时间时长	连续变量	0~729
		平均乘机时间间隔	连续变量	0~720
		第2年的乘机次数比率	连续变量	0~1
		第1年的乘机次数比率	连续变量	0~1
		平均折扣率	连续变量	0.11~1.5
		第1年乘机次数	连续变量	0~118
		第2年乘机次数	连续变量	0~111
自变量	用户积分数据	第1年里程积分	连续变量	0~186104
		第2年里程积分	连续变量	0~166331
		观测窗口总精英积分	连续变量	0~74460
		观测窗口中其他积分	连续变量	0~984938
		非乘机积分总和	连续变量	0~984938
		总累计积分	连续变量	0~502544
		非乘机的积分变动次数	连续变量	0~140
		总基本积分	连续变量	0~286164
		观测窗口中第1年其他积分	连续变量	0~600000
		观测窗口中第2年其他积分	连续变量	0~728282
		积分兑换次数	连续变量	0~46

4.1.3 流失客户特征分析

将全部航空公司客户数据进行 logistic 回归分析，通过回归结果显示客户流失与否和选取的 30 个变量中的 18 个呈现出高度相关，结果见下表 4.2。

表 4.2 逻辑回归系数

	Estimate	Std.Error	z value	Pr(>|z|)	显著性
(Intercept)	7.901e+01	6.860e+00	11.518	< 2e-16	***
L	-5.336e-03	5.149e-04	-10.364	< 2e-16	***
GENDER	8.717e-02	2.768e-02	3.149	0.001638	
FFP_TIER	-3.102e-01	5.518e-02	-5.621	1.90e-08	**
WORK_COUNTRY	-2.349e-01	4.495e-02	-5.225	1.74e-07	
age	-2.274e-03	1.285e-03	-1.770	0.076688	.
FLIGHT_COUNT	-3.628e-01	1.832e-02	-19.808	< 2e-16	***
BASE_POINTS_SUM	-3.687e-05	1.709e-05	-2.157	0.031008	*
EXPENSE_SUM_YR_1	-2.709e-05	6.316e-06	-4.289	1.79e-05	***
EXPENSE_SUM_YR_2	-5.486e-05	9.167e-06	-5.985	2.17e-09	***
SEG_KM_SUM	-5.589e-05	5.120e-06	-10.917	< 2e-16	***
WEIGHTED_SEG_KM	7.468e-05	1.129e-05	6.614	3.74e-11	***
AVG_FLIGHT_COUNT	-1.072e	1.186e-01	-9.036	< 2e-16	***
AVG_BASE_POINTS_SUM	1.209e-04	9.637e-05	1.254	0.209689	
DAYS_FROM_BEGIN_TO_FIRST	2.632e-03	1.219e-04	21.594	< 2e-16	***
AVG_FLIGHT_INTERVAL	2.676e-02	3.652e-04	73.279	< 2e-16	***
ADD_POINTS_SUM_YR_1	5.582e-06	3.853e-06	1.449	0.147414	
ADD_POINTS_SUM_YR_2	-5.102e-06	4.190e-06	-1.218	0.223340	
EXCHANGE_COUNT	1.025e-01	1.754e-02	5.847	5.00e-09	***
avg_discount	-2.729e-01	9.380e-02	-2.909	0.003621	**
P1Y_Flight_Count	1.130e-01	1.106e-02	10.217	< 2e-16	***
L1Y_Flight_Count	1.138e-01	1.524e-02	5.847	4.05e-01	
P1Y_BASE_POINTS_SUM	-1.897e-05	1.026e-05	-1.848	0.064551	.
L1Y_BASE_POINTS_SUM	-1.095e-05	1.356e-05	-1.359	0.053271	
ELITE_POINTS_SUM	3.260e-04	2.234e-05	14.593	< 2e-16	***
ADD_POINTS_SUM	-1.897e-05	1.026e-05	-1.848	0.064551	
Eli_Add_Point_Sum	-1.897e-05	1.026e-05	-1.848	0.064551	
Points_Sum	-1.897e-05	1.026e-05	-1.848	0.064551	
Ration_L1Y_Flight_Count	-7.586e-01	6.850e+00	-11.074	< 2e-16	***
Ration_P1Y_Flight_Count	-7.262e-01	6.850e+00	-10.602	< 2e-16	***
Point_Chg_NotFlight	-7.464e-03	2.253e-03	-3.314	0.070921	***

注：Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’

从客户基本资料信息来看，航空公司客户会员卡等级和入会时长对客户是否流失的影响是比较显著的，在其他自变量不变的情况下，某个自变量的变化可以反映出对客户流失影响的方向和程度。会员卡等级反映出客户在航空公司的消费频数，即等级越高乘坐本公司航班次数越多或消费金额越高，会员卡等级的系数小于零，说明客户的等级越高越不容易流失，这与我们平常的经验相一致。客户入会时间长的系数也是小于零的，说明客户入会时间越久越不易流失。

从用户飞行信息来看，飞行次数、总票价、乘坐里程、第一次乘机时间至入会时间时长、平均乘机间隔、平均折扣率与客户流失均呈现出强相关性。第一次乘机时间至入会时间时长与客户流失正相关，平均乘机间隔与客户流失正相关，客户在加入航空公司的会员后，却一直都没有乘坐本公司航班，时间越长越会增加客户流失的可能性，同样的，客户的平均乘机间隔越长，客户流失的可能性也越大。而飞行次数、总票价、乘坐里程、平均折扣率与客户流失负相关，飞行次数越多、总票价越高、客户乘坐本公司航班的里程越多、平均折扣率越高，客户越不易流失。

从用户积分信息来看，总基本积分和非基本积分兑换次数、第 1 年里程积分占最近两年积分比例的系数为负，总基本积分越高、非基本积分兑换次数越多、第 1 年里程积分占最近两年积分比例越高的客户越不易流失。总精英积分越高的客户越容易流失。

4.1.4 参数选取及预测结果分析

前面使用 logistic 回归对流失客户的特征进行了分析，接下来使用 logistic 回归模型对流失客户进行识别。将数据集划分为训练集（数据集的 80%）、测试集（数据集的 20%）。使用训练集训练逻辑回归模型，并利用所得的模型计算其在测试集上的分类结果，最后计算分类的准确率和 *AUC* 。

首先使用训练数据 X_train 和 y_train，由易到难地建立多个不同的逻辑回归模型。在第一个逻辑回归模型中，在训练模型时不使用正则化项，所得的模型记为 M1，之后使用 predict 函数得到训练集的测试结果。在第二个逻辑回归模型中，仅使用 L_1 范数作为正则化项，将正则化参数设置为 0.01，然后改变正则化参数的数值，得到不同参数下逻辑回归模型识别流失客户的召回率（覆盖面的度量，度量有多个正例被分为正例，召回率越大模型效果越好），使用交叉检验选出最优的模型。表 4.3 给出了不同正则化参数下的预测客户流失模型在训练集上的召回率，表格显示，当正则化参数取值为 10 时，召回率最大。

表 4.3 不同正则化参数的预测客户流失模型训练集召回率

正则化参数	0.01	0.1	1	10	100
召回率（%）	74.28	74.52	74.62	74.64	74.64

另外，不同的切割率 P(0<P<1)也会影响模型的性能。为了选取出使模型准确率更高是的切割率，本文选取不同的切割率(0.1～0.9)，将测试集的数据纳入正则化参数为 10 的逻辑回归识别航空公司流失客户模型，得出九个模型的分类结果，这里给出切割率为 0.4、0.5 时的分类结果，当切割率取值为其它数值时，准确率较低，故此处未给出分类结果。

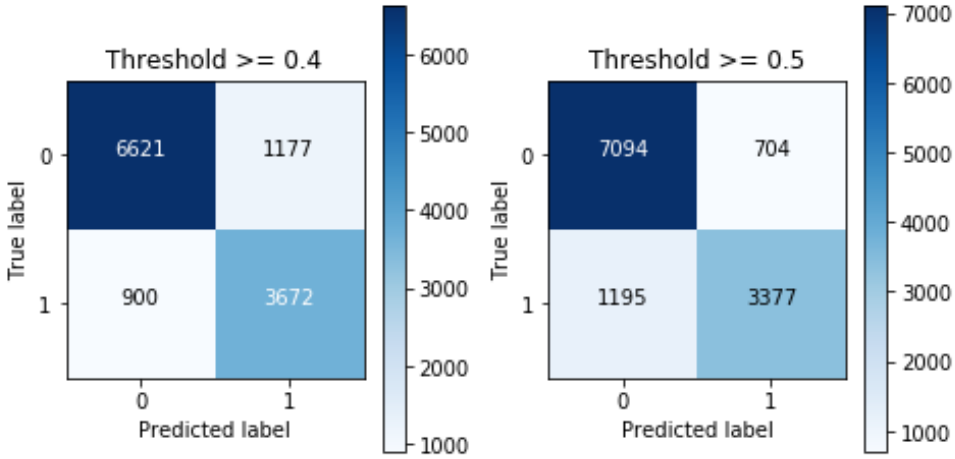


图 4.2 切割率分别为 0.4、0.5 时模型训练集分类结果

根据本文研究的实际问题，当航空公司把流失客户被错误地分为非流失客户时，航空公司就会忽视对这些客户的管理，会导致很大的损失，基于这种分析，选取 0.4 的切割率较 0.5 好。最终模型的正则化参数选取 0.01，切割率选取 0.5，得到的测试集分类结果及 *ROC* 曲线见下图。

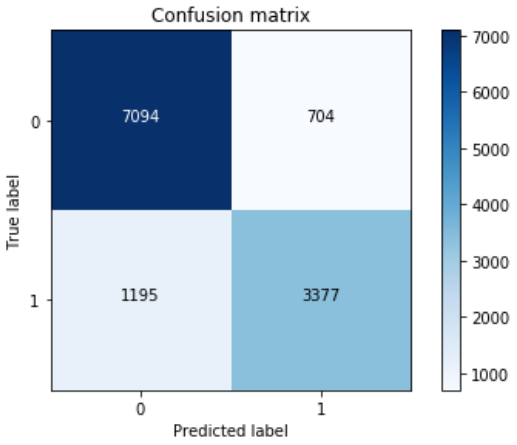


图 4.3 *logistic* 识别流失客户模型测试集分类结果

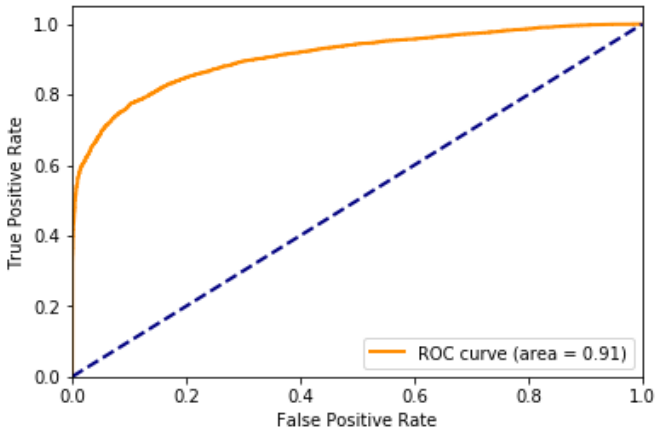


图 4.4 *logistic* 识别航空公司流失客户模型在测试集上的 *ROC* 曲线

图 4.3 显示了正则化参数的数值为 10，切割率为 0.5 时，*logistic* 回归模型预测测试集流失客户结果。其中，73.86%的流失客户被准确地预测出来，90.97%的非流失客户被准确地预测出来，流失客户的正确预测率表示阳性预测值，非流失客户的正确预测率表示阴性预测值。真非流失客户的假阳性比率（即真实情况下是非流失客户而在预测模型中却被错误地判断为流失客户，占真实情况下非流失客户的比例）为 9.03%，真流失客户的假阴性比率（即真实情况下是流失客户而在预测模型中却被错误地判断为非流失客户，占真实情况下流失客户的比例）为 26.14%。流失客户预测中的假阳性率（即识别模型判断出的所有流失客户中，真实情况是非流失客户，占所有预测出的流失客户的比例）为 17.25%。非流失客户预测中的假阴性率（即预测模型判断出的所有非流失客户中，真实情况是流失客户，占所有预测出的非流失客户的比例）为 14.42%。模型的总体准确度为 84.65%。由图 4.4 可以看出 *AUC* 值为 0.91。

4.2 航空公司客户流失的随机森林模型

上一小节介绍了航空公司客户流失的 logistic 回归模型，最终模型的总体识别准确率为 84.65%，分类精度并不高。而对于分类问题，随机森林模型在各种竞赛中被广泛使用，本节就使用集成学习中的随机森林模型来建立客户流失预测模型，以提高分类精度。本节主要内容如下：使用训练集建立合适的航空公司客户流失的随机森林模型，在预测集上分析预测结果，最后使用随机森林研究变量的重要性。

4.2.1 随机森林介绍

集成学习中，通过构建一系列的模型（基学习器），使用不同策略，将这些基学习器聚合起来，以此来提高模型性能。根据基学习器的生成策略，可将集成学习的方法可以分为两类：

（1）并行方法（ *parallel method* ），以 *bagging* 为主要代表；

（2）顺序方法（ *sequential method* ），以 *boosting* 为主要代表。

随机森林使用了 *bagging* 的基本思想来训练一系列的决策树[32]。但是随机森林又根据决策树的特点做了很多改进，使得所构建的决策树尽量没有相关性，从而显著提高最终的性能。在很多实际问题中，随机森林都能取得很好的效果，同时由于其控制参数易于选择，因此使得随机森林成为实际应用中非常受欢迎的算法。

而对于决策树，它能够有效地获取多个变量之间的相互关系。同时，如果决策树足够深，那么它能够有效地降低模型的偏差。较深的决策树虽然能够降低模

型的偏差，但是很容易导致过拟合。对于不同的数据集，得到的决策树可以存在较大差异。换言之，决策树的方差较大，因此是使用*bagging*的好对象。

（1）训练随机森林的基本流程

对于随机森林，其基本思想是利用*bagging*构建很多决策树。为了提高*bagging*的效果，需要构建尽量独立的基学习器。虽然每个基学习器的性能表现比较弱，但是若干个学习器都是从不同角度犯的错，那么将它们聚合起来后，性能也许会有大提升。在实际的问题中，给定训练集，建立多个相对独立的基学习器有两种选择：（1）使用不同的训练集，使用相同的学习算法训练出许多基学习器；（2）当训练集一样或者训练集很相近时，使用不同的学习算法，来训练出不同的基学习器。

取样：从n个样本的数据集中取出n个样本。假设样本共有n个，使用*bootstrap*取样方法，取出的n个样本中肯定会有重复的样本出现，同时也会出现原来的样本中有未被取出的样本的情况，除非取样时出现极端情况（*bootstrap*取样时，取出了所有的原样本，这样得到的是整个的原数据集）。注意，在*bagging*中，对于原始的训练集，使用*bootstrap*取样m次，选取出m个样本集。在m个样本集，分别构建基学习器。这样就能得到m个不同的学习器。

模型聚合：对很多个模型进行聚合工作时，选用的方法很简单。在分类问题中，采用投票（*voting*）的方法，m个基学习器分类将会出现m个结果，m个结果中出现次数最多的那个结果将会被当作最终的分类结果；而在回归问题当中，m个基学习器将会出现m个输出结果，取m个结果的平均值作为最终的结果。另外，对于多分类的问题，也可以对其进行处理，只要能训练出处理这类问题的基学习器就可以。

在随机森林中，要构建出一系列的决策树。在构建每个决策树时，首先使用*bootstrap*取样得到一个样本数为n的可重复样本。然后利用这个样本集，构建一棵决策树。在构建决策树时先随机取出d_1个变量，再选取出最优的变量。最后将所有模型的输出取平均（回归问题）或者取出最多的类别（分类问题）作为最终的输出。

（2）利用随机森林估计变量的重要性

在实际使用机器学习算法时，通常数据包含噪声或者冗余数据。因此，如果模型能够直接告诉我们各个变量的重要性，就可以剔除这些噪声和冗余数据。不仅可以改善模型的性能，而且可以降低计算复杂度。

在随机森林中，通常可以使用*OOB*样本来确定每个变量的重要性。在考虑第j棵树T_j时，将所有没有包含在第j次*bootstrap*取样集S_j中的*OOB*样本集记为O_j，将其当做检验集（*validationset*）。当考虑第k个变量的重要性时，把检验

集中样本的第k个变量的值随机打乱，得到新数据集O'_{jk}。分别计算T_j在O_j和O'_{jk}上的性能（如准确率）P_j 和P'_{jk}。将P_j与P'_{jk}的差作为第k个变量的重要性。在随机森林中，有m棵树，将所有m棵树P_j与P'_{jk}的差平均起来，就可以作为第k个变量的重要性的度量。以分类问题为例，假设选定性能度量为准确率，一般来说，$P_j \geq P'_{jk}$，第k个变量重要性为：

$$IM(v_k)=\frac{1}{m}\sum_{j=1}^{m}(P_j-P'_{jk}) \tag{4.10}$$

这种通过随机排列变量值来估计变量重要性的方式，目前也有了很多应用。在微软公司最新推出的 *Azure Machine Learning* 中就有相应的模块来通过此法估计变量的重要性。

（3）随机森林的特点

- 具有极好的准确率
- 对于很大的数据集，也能够有效地运行
- 当样本数据具有高维特征时，也能够对其进行处理，不需对其进行降维
- 能够很好的评估每一个特征在分类问题上的重要性
- 在随机森林生成的过程中，是一种能够取得到内部生成误差的无偏估计

4.2.2 参数选取与预测结果分析

随机森林预测航空公司客户流失模型中参数有两种：一种是模型中树的个数，另一种是控制每棵树大小的参数。首先使用训练集构建随机森林模型，在这个模型中，全部采用默认的参数值 n_estimators=10，min_samples_split=2，min_samples_leaf=1，为了获得较为精确的准确率，本文采用交叉检验的方法，得到准确率为 92.02%。固定树的大小（min_samples_split=2，min_samples_leaf=1）后向该随机森林模型中不断添加决策树的数目，随着树的数目增加，训练集的正确率变化如下图所示：

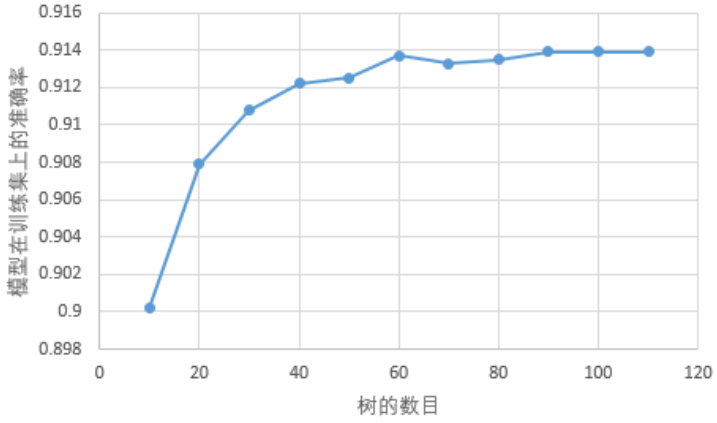


图 4.5 不同树的数目下随机森林识别流失客户模型在训练集上的正确率

根据图 4.5 可知，随着树数目的增加，模型在训练集上的准确率先增加，树的数目为 90 时，训练集的准确率增加到 91.39%，继续增大树的个数后，准确率变化不大，所以选取参数 n_estimators=90 。然后固定随机森林中树的数目 90，改变控制每棵树大小的参数最小中间节点样本数和最小叶子节点样本数，改变参数 min_samples_split ， min_samples_leaf 后，模型在训练集上的准确率均出现下降的结果，最终预测航空公司客户流失的随机森林模型的参数选取为：树的数目 90，每棵树中中间结点最小样本数为 2、最小叶子节点样本数为 1，最后得出训练集和测试集的分类结果如下图所示。

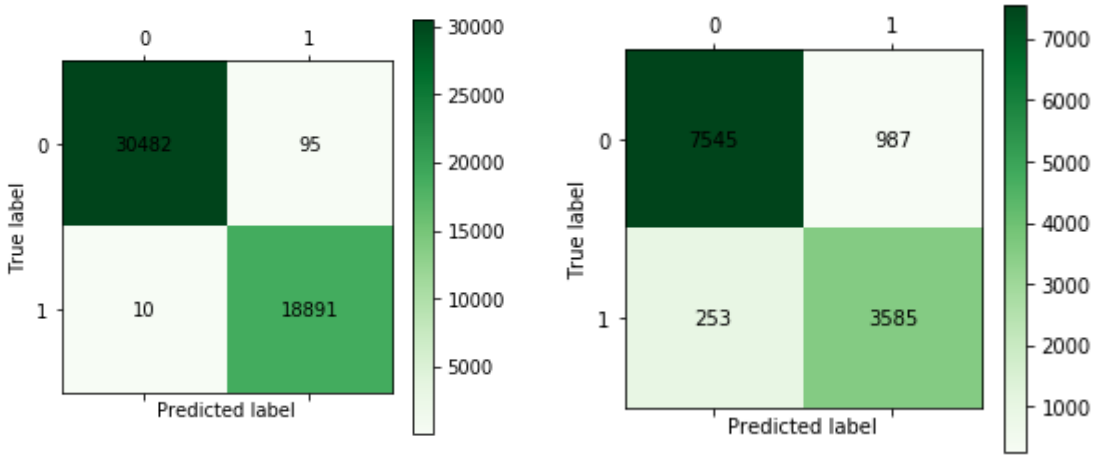


图 4.6 随机森林识别流失客户模型在训练集和测试集分类结果

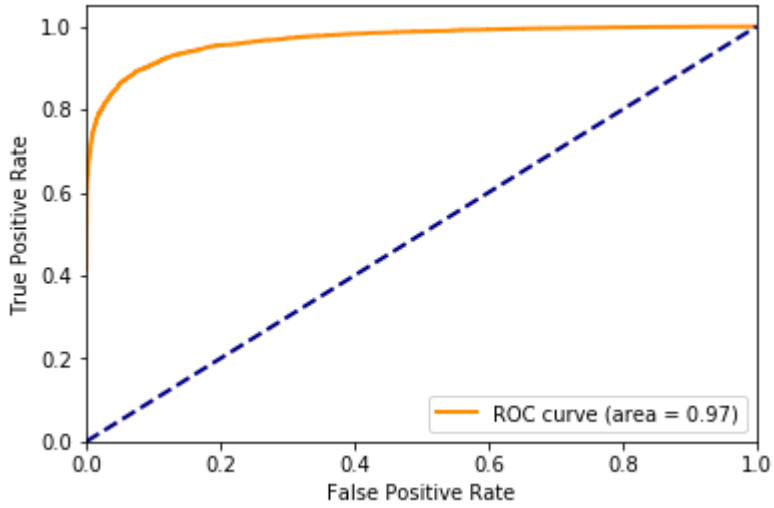


图 4.7 随机森林识别航空公司流失客户模型在测试集上的 ROC 曲线

上图显示，树的数目为 90 时，随机森林预测测试集流失客户结果。其中，93.41%的流失客户被准确地预测出来，88.43%的非流失客户被准确地预测出来，流失客户的正确预测率表示阳性预测值，非流失客户的正确预测率表示阴性预测值。真非流失客户的假阳性比率（即真实情况下是非流失客户而在预测模型中却被错误地判断为流失客户，占真实情况下非流失客户的比例）为 11.57 %，真流

失客户的假阴性比率（即真实情况下是流失客户而在预测模型中却被错误地判断为非流失客户，占真实情况下流失客户的比例）为 6.59%。流失客户预测中的假阳性率（即识别模型判断出的所有流失客户中，真实情况是非流失客户，占所有预测出的流失客户的比例）为 21.59%。非流失客户预测中的假阴性率（即预测模型判断出的所有非流失客户中，真实情况是流失客户，占所有预测出的非流失客户的比例）为 3.24%。模型的总体预测准确度为 89.98%。由图 4.7 可以看出 *AUC* 值为 0.97。

4.2.3 随机森林预测流失客户模型变量重要性

使用 SelectKBest 函数得到该模型中变量的重要程度，得到的变量重要程度如下图所示。

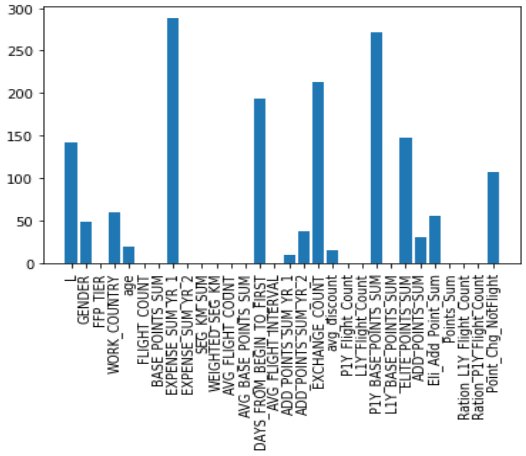


图 4.8 随机森林识别航空公司流失客户模型中变量的重要程度

根据上图显示结果可知，作为识别出是否是流失客户的前五个重要影响特征分别是：

表 4.4 五个重要影响特征

重要程度	特征名称	特征意义
1	EXPENSE_SUM_YR_1	第 1 年票价总收入
2	P1Y_BASE_POINTS_SUM	第 1 年里程积分
3	EXCHANGE_COUNT	积分兑换次数
4	DAYS_FROM_BEGIN_TO_FIRST	观察窗口内第一次乘机时间至入会时间时长
5	ELITE_POINTS_SUM	观测窗口总精英积分

4.3 航空公司客户流失的梯度提升树模型

上一小节中介绍的航空公司客户的随机森林模型的总体预测准确度为

89.98%，比 logistic 模型的准确度只提高了 5.33%，预测准确度并没有很大的提高，梯度提升树模型作为 boosting 的典型算法，具有可解释强、容易并行等优点，本小节建立航空公司客户的梯度提升树模型，以进一步提高预测准确度。本节主要内容如下：使用训练集建立合适的航空公司客户流失的梯度提升树模型，在预测集上分析预测结果，最后使用梯度提升树研究变量的重要性。

4.3.1 梯度提升树介绍

在梯度提升算法中，每次根据已有的分类情况，构建出新的目标值。在梯度提升树中，可以使用不同的损失函数，通过给出的训练集 $\{(x_1, y_1),(x_2, y_2), \cdots,(x_n, y_n)\}$，最小化模型 $f(x)$ 对应的损失函数 $L(y,f)$ 来求解最优的模型：

$$\min_f L(y,f)=\sum_{i=1}^{n} L(y_i, f(x_i)) \tag{4.11}$$

假设 $f(x)$ 表示一系列决策树的和：

$$f(x)=\sum_{j=1}^{m} h_j(x) \tag{4.12}$$

把第 j 棵树的对应函数记为 $h_j(x)$。把第 $J(J \le m)$ 步得到的模型 $f_J(x)$ 记作 $\sum_{j=1}^{J} h_j(x)$。

在梯度提升树算法中，按照顺序建立模型 $h_j(x)$，其核心思想是：在第 $j(j \le m)$ 步构建 $h_j(x)$ 模型的时候，用负梯度 $-[\frac{\partial L(y_i,f)}{\partial f}]_{f=f_{j-1}}$ 当作 x_i 的新目标值。在不同的问题中，需要选择不同的损失函数，并且对不同的损失函数，可以计算相应的负梯度，并作为新的目标值来训练新的学习器，而在训练新学习器时是利用训练集数据来计算的导数，最小化损失函数也是基于训练集的数据。然而，在实际问题中，既希望模型在训练集上取得好的效果，也希望模型在测试集上也能表现优秀，也就是要避免模型在测试集出现过拟合，一般可以采取控制树的个数拟合负导数。

梯度提升算法的步骤：

1. 得到初始函数 $f_0(x)=c$，$c=\arg\min_c \sum_{i=1}^{n} L(y_i,c)$

2. *for* $j=1:m$

2.1 计算负梯度：

$$r_{ij}=-[\frac{\partial L(y_i,f)}{\partial f}]_{f=f_{j-1}} \tag{4.13}$$

2.2 把 $r=\left[r_{1j}, r_{2j}, \ldots, r_{nj}\right]^T$ 当做新目标，在训练集上构建新模型

2.3 求解步长 s_j：

$$s_j = \arg\min_{s>0} \sum_{i=1}^{n} L(y_i, f_{i-1}(x_i) + sh_j(x_i)) \tag{4.14}$$

2.4 更新函数 f：

$$f_j(x) = f_{j-1}(x) + s_j h_j(x) \tag{4.15}$$

3 输出模型：

$$f(x) = f_m(x) \tag{4.16}$$

在实际使用梯度提升树时，可以使用以下技巧来提高算法的性能，并有效地避免过拟合现象。

- 控制树的个数；
- 控制每棵树的大小；
- 控制学习率；
- 子取样。

1.学习率和树的数目

在梯度提升树中，可以构建许多棵树。从直观上讲，最开始构造的决策树更多地描述了最后所得模型的主要框架；而后面的决策树更多地是考虑那些难分类的样本。因此，可以认为应该给前面的决策树更大的权重。而随着更多的决策树的建立，应该逐渐降低其权重。具体来说，在建立第 j 个模型时，使用如下的公式来更新模型：

$$f_j(x) = f_{j-1}(x) + vh_j(x) \tag{4.17}$$

参数 v 满足 $0 < v < 1$，是新的决策树 $h_j(x)$ 的权重。如果把这个公式和数值最优化中的更新公式相比较，可以把 v 看成是沿着负梯度方向移动的步长，因此，v 也可以认为是控制了 *boosting* 算法的学习率。另外，可以将学习率与正则化相联系。

在实际使用梯度提升树算法时，树的数目和学习率是相互依赖的。一般而言，v 较小时，要以较慢的速度去拟合目标函数，需要构建较多的决策树；v 较大时，可以构建较少的决策树。在实际中，较小的学习率一般能够得到更好的模型。举一个简单的例子，将 v 设置为 0.005，在 *OOB* 样本中的表现，会显著强于将 v 设置为 0.05 时的情况。但是为了获得较好的性能，较小的 v 值将需要更多的树，但这需要更多的存储空间和计算时间。如果将 v 从 0.05 降到 0.005，基本上需要多近十倍的计算时间。

2.树的大小

控制树的大小可以通过控制树的深度或者叶结点的数目来实现。由于在梯度提升算法中我构建了多棵决策树，因此一般倾向于每棵决策树的复杂度不宜太高。例如，每个决策树的叶结点的数量可以被设置为一个恒定值，或者每棵决策树的深度是恒定值。而这些都可以作为梯度提升算法的参数，可根据实际的数据予以

调节。

3.子取样

子取样的思想与*bootstrap*取样类似。在随机森林中，通过*bootstrap*取样显著降低了生成的决策树之间的相关性。在梯度提升算法中，也可以采用类似的方法。具体来说，在使用子取样时，从所有的样本中随机取出一部分（但不是重复取样）的样本。这样，每次构建新的决策树时，只通过子取样得到一部分训练样本来构建新的模型。通过使用子取样，能够在构建决策树的时候引入一些随机性，从而能够增强各个模型的多样性，进而提高聚合之后的模型的性能。

4.3.2 参数选取与预测结果分析

在训练集上选用默认参数建立梯度提升树模型，准确率为 97.75%。使用交叉验证选取最优参数后，模型在训练集上的正确率由 97.75%提高到 99.9%。最优参数见下表。

表 4.5 模型参数选取

参数名称	参数意义	数值
learning_rate	学习率	0.3
n_estimators	树的数目	120
max_depth	树的最大深度	5
min_samples_split	分支最小样本量	2
min_samples_leaf	叶节点最小样本量	1
max_features	最大特征量	13

使用以上参数对测试集建立梯度提升树模型来预测航空公司客户流失，预测结果如下：

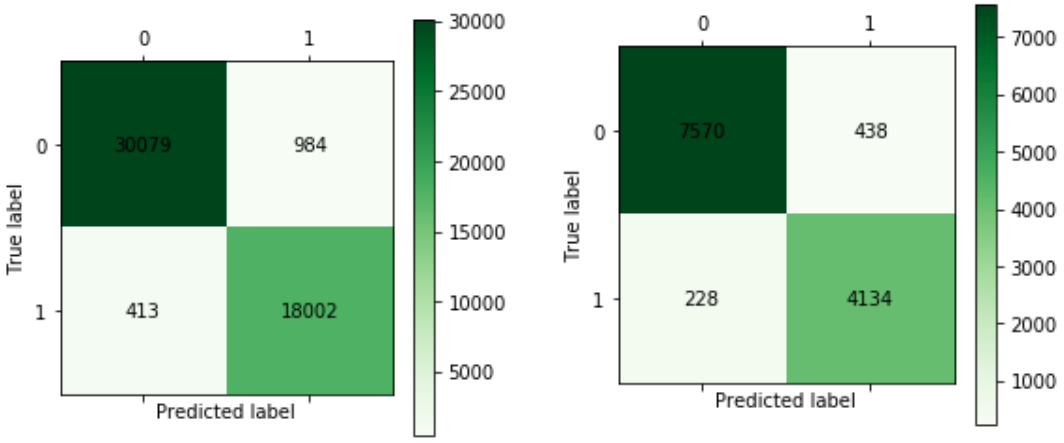


图 4.9 梯度提升树识别流失客户模型在训练集和测试集的分类结果

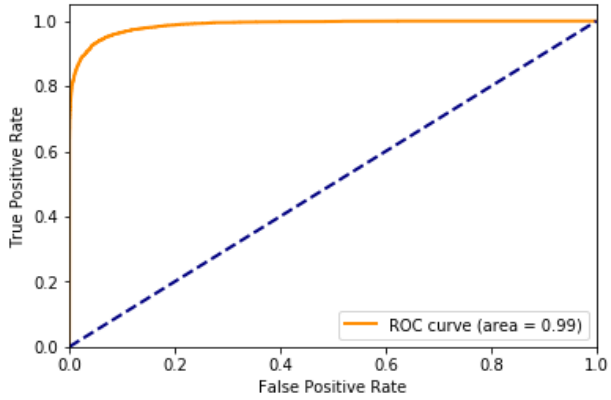


图 4.10 梯度提升树识别航空公司流失客户模型在测试集上的 *ROC* 曲线

梯度提升树识别测试集流失客户结果显示，94.78%的流失客户被准确地识别出来，94.93%的非流失客户被准确地识别出来，流失客户的正确识别率表示阳性预测值，非流失客户的正确识别率表示阴性预测值。真非流失客户的假阳性比率（即真实情况下是非流失客户而在识别模型中却被错误地判断为流失客户，占真实情况下非流失客户的比例）为 5.07%，真流失客户的假阴性比率（即真实情况下是流失客户而在识别模型中却被错误地判断为非流失客户，占真实情况下流失客户的比例）为 5.22%。流失客户识别中的假阳性率（即识别模型判断出的所有流失客户中，真实情况是非流失客户，占所有识别出的流失客户的比例）为 9.58%。非流失客户识别中的假阴性率（即识别模型判断出的所有非流失客户中，真实情况是流失客户，占所有识别出的非流失客户的比例）为 2.92%。模型的总体识别准确度为 94.62%。

4.3.3 梯度提升树预测流失客户模型变量重要性

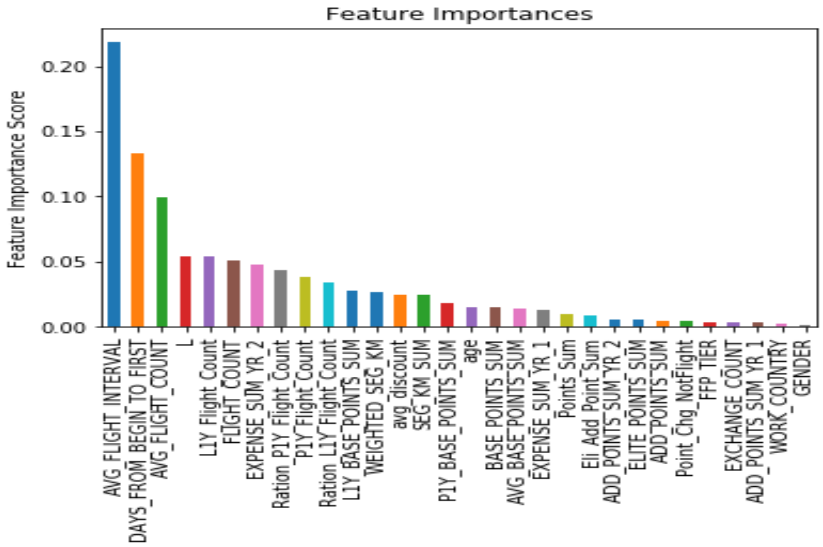


图 4.11 梯度提升树模型识别航空公司流失客户模型中变量的重要程度

根据上图显示结果可知，作为识别出是否为流失客户的前五个重要影响特征

分别是：

表 4.6 五个重要影响特征

重要程度	特征名称	特征意义
1	AVG_FLIGHT_INTERVAL	平均乘机间隔
2	DAYS_FROM_BEGIN_TO_FIRST	观察窗口内第一次乘机时间至入会时间时长
3	AVG_FLIGHT_COUNT	平均乘机次数
4	L	入会时长
5	L1Y_FLIGHT_COUNTER	第一年飞行次数

4.4 三种航空公司客户流失预测模型性能的对比分析

用于计算分类算法性能的各种评价标准有准确率、混淆矩阵、精确率、召回率、*F*1度量、*ROC* 曲线和 *AUC* 等。

1.准确率

准确率（*accuracy*）的定义为：

$$准确率=\frac{正确分类的样本数}{总样本数} \tag{4.18}$$

根据定义，准确率的取值在0和1之间，越大说明分类结果越好。

2. 混淆矩阵

在流失预测的例子中，其中错误的分类有两种：非流失客户被判定为流失；流失客户被判定为非流失。

显然，我们对第二种错误更感兴趣，因为它所对应的代价更高，更应该避免。在这种情况下，可以考虑使用混淆矩阵来比较算法。

在两类分类问题中，2×2 的混淆矩阵的各个元素有专门的名字。其中 *True Positive*（*TP*）表示实际是正类，同时也被判定为正类，*False Positive*(*FP*) 表示实际是负类，但是被判定为正类；*False Negative*（*FN*）表示实际是正类，但是被判定为负类；*True Negative*（*TN*）表示实际是负类，同时也被判定为负类。

表 4.7 混淆矩阵

预测状态 / 实际状态	1	0	合计
1	TP	FN	P
0	FP	TN	N
合计	P'	N'	

定义 *False Negative Rate*（*FNR*）为正类样本中被错误分类的比例：

$$FNR = \frac{FN}{FN + TP} \tag{4.19}$$

准确率（*accuracy*）：

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.20}$$

3.精确率、召回率和F1度量

在本文的航空公司客户流失预测中，我们主要关注流失的客户，可以将流失客户定义为正类，而非流失客户定义为负类。对于这样的问题，常用的评价标准有精确率（*precision*）、召回率（*recall*）和*F*1度量（$F1-Measure$），其具体定义如下：

$$precision = \frac{TP}{TP + FP} \tag{4.21}$$

$$recall = \frac{TP}{TP + FN} \tag{4.22}$$

$$F1 = \frac{2precision \times recall}{precision + recall} = \frac{2TP}{2TP + FP + FN} \tag{4.23}$$

4. *ROC* 曲线和 *AUC*

在 *ROC* 曲线中，横坐标是 *False Positive Rate*（*FPR*），纵坐标为 *True Positive Rate*（*TPR*）。在 *ROC* 曲线中，要着重考虑以下四个点，第一个点是 $(0,0)$，即 $FPR = TPR = 0$，也即 $FP = TP = 0$。此时分类阈值极大，分类器将所有的样本都判定为负样本了。第二个点是 $(0,1)$，即 $FPR = 0$，$TPR = 1$，也即 $FP = FN = 0$。此时分类器将所有的样本都正确分类了。第三个点是 $(1, 0)$，即 $FPR = 1$，$TPR = 0$，也即 $TP = TN = 0$。此时分类器将所有的样本都错误分类了。第四个点是 $(1,1)$，即 $FPR = 1$，$TPR = 1$，也即 $TN = FN = 0$。此时分类阈值极小，分类器将所有的样本都判定为正类了。*ROC* 曲线越接近左上角（点(0,1)），分类器的性能越好。实际中经常使用 *ROC* 曲线下的面积来衡量算法的好坏，称为 *AUC*（area under ROC curve）。可以看出 $0 < AUC < 1$。如果分类器 f 的输出对于任意的负类样本 x_- 和正类样本 x_+ 满足 $f(x_-) < f(x_+)$，则该分类器的 *ROC* 曲线经过点 $(0,1)$，对应的 *AUC* 值为 1。如果一个分类器随机猜测样本对应的类别，其对应的 *AUC* 为 0.5。一般来说，分类器的 *AUC* 介于0.5~1之间。

除了 *ROC* 曲线下的面积，*AUC* 还有更加直观的解释。可以严格地证明 *AUC* 等于任意选择的一对负类样本 x_- 和正类样本 x_+ 满足 $f(x_-) < f(x_+)$ 的概率，即

$$AUC = P(f(x_+) > f(x_-)) \tag{4.24}$$

AUC 的值越大，分类器的性能越好。在实际中，有些算法会直接利用上式来优化 *AUC* 值。*ROC* 曲线的另外一个良好的性质是：即便在类不平衡的时候，

它都是一个良好的算法评价指标。而在这种情况下，常用的准确率等评价 指标容易受到样本数较多的类的影响，不能很好地反映算法实际的好坏。

4.4.1 模型评估指标对比分析

定义以下模型的预测结果，客户是否流失=1 称作是“正”（positive）。得到的结果：

表 4.8 模型评估指标

评价	logistic	随机森林	梯度提升树
准确率	84.65%	89.98%	94.62%
精确率	82.75%	78.41%	90.42%
召回率	73.86%	93.41%	94.77%
FNR	26.14%	6.59%	5.23%

根据本文研究的问题是航空公司客户流失模型分类结果会出现两种错误：一种是流失客户被错误地分为非流失客户，另一种是非流失客户被错误得分类为流失客户。对于后一种错误，客户真实情况是非流失客户，尽管被错认为流失客户，航空公司对其采取一定措施，航空公司也不会也什么大的损失；而前一种错误，客户真实情况是流失客户，却被误分为非流失客户，那么航空公司就会忽视对这些的管理，会导致较大的损失，这种错误是最不应该犯的，所以评价指标 *FNR*（正类样本中被错误分类的比例）越小越好。对比三个预测模型，梯度提升树算法模型的 *FNR* 最小且准确率也比另两个模型要高。

4.4.2 模型 ROC 曲线对比分析

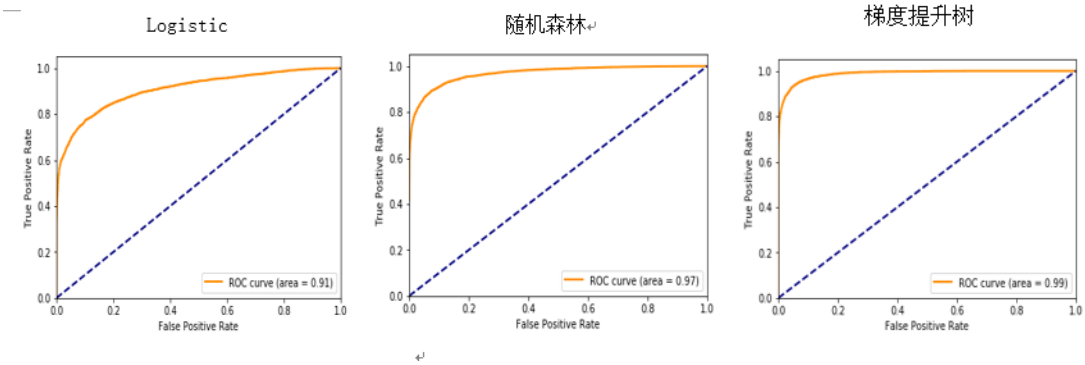


图 4.12 三个模型的 ROC 曲线

可以看出，这三个模型的真负率（特指度）分别为 90.97%、93.41%、94.77%，真正率（灵敏度）分别为 73.86%、88.43%、94.5%，对于测试集的准确率分别为 84.65%、89.98、94.62%， *AUC* 值等于 0.91、0.97、0.99，对于这三种航空公司客户流失模型，梯度提升树算法效果是最好的，且根据模型预测的结果可知无过拟合现象。

5 结论和展望

5.1 主要工作总结

本文对航空公司客户进行了分类，对预测航空公司客户流失模型做了探讨。在理论研究方面，本文介绍了客户分类和预测客户流失的相关研究。对航空公司的会员客户数据做了聚类分析，依客户行为特征将客户划分为五类，对各类客户的特征做了论证分析。结果显示重要保持客户乘坐次数或乘坐里程均较高，平均折扣率较高,最近乘坐过本公司航班，这类客户的价值很高，航空公司应该优先将资源投放到这类客户身上。使用层次分析法确定客户行为指标权重，得出每类客户的价值量,弥补分类的不足。对航空公司的流失客户识别模型进行深入探讨，选用 logistic、随机森林和梯度提升树模型对航空公司客户进行流失预测。首先对流失客户做出定义：最后一次乘机时间至观测窗口末端时长大于等于观测窗口内最大乘机间隔。使用全部客户数据进行 logistic 回归，用回归系数对流失客户的特征进行分析。随机将 80%的数据分为训练集训练分类器，构建客户流失预测的 logistic 回归模型、随机森林模型和梯度提升树模型，并从随机森林模型和梯度提升树模型中得出影响客户流失的五个重要变量。最后，从模型指标评估指数方面对三个流失预测模型进行对比分析，这三个对于流失客户的识别模型效果最好的是梯度提升树模型，且根据模型预测的结果可知无过拟合现象。

5.2 研究展望

在对各类客户进行特征分析时进一步对其特征进行内部机制分析。在建立流失预测模型时，对全体客户进行的流失预测，对于预测出的可能流失的客户，并不是都值得挽留的，可以结合客户的价值判断客户是否值得挽留。

参考文献

[1]许力梅. 基于关联规则的决策树算法改进及应用[D].华南理工大学,2011.

[2]李琪,崔睿.保险业在线客户细分与忠诚度研究——基于 SOM 神经网络模型[J].经济经纬,2012(06):62-66.

[3]郜丽君,胡如夫,赵韩,陈曹维.基于 SOM&SVM 组合分类器的客户细分方法实证研究[J].数学的实践与认识,2012,42(11):139-146.

[4]王园.证券业客户细分模型构建及实证研究[J].上海管理科学,2012,34(02):30-35.

[5]宋中山,周腾,周晶平.一种改进的蚁群聚类算法在客户细分中的应用[J].中南民族大学学报(自然科学版),2013,32(03):77-81.

[6]曾小青,徐秦,张丹,林大瀚.基于消费数据挖掘的多指标客户细分新方法[J].计算机应用研究,2013,30(10):2944-2947.

[7]颜书云.信用卡客户细分实证研究[J].中国管理信息化,2013,16(17):34-35.

[8]张劲松,江波.基于 C4.5 算法的民航客户价值细分研究[J].西安航空学院学报,2014,32(05):75-78+96.

[9]王光辉,张晓光,赵艳芹.基于遗传算法和 BP 神经网络的多维客户行为细分模型的研究[J].齐齐哈尔大学学报(自然科学版),2014,30(04):37-39.

[10]宋才华,蓝源娟,王永才,翟鸿荣,李滨涛.综合价值评估法在电力企业客户细分中的应用[J].电子设计工程,2014,22(12):111-113+116.

[11]胡晓雪,赵嵩正,吴楠.基于 SOM-DB-PAM 混合聚类算法的电力客户细分[J].计算机工程,2015,41(10):295-301+308.

[12]潘俊,王瑞琴.基于选择性聚类集成的客户细分[J].计算机集成制造系统,2015,21(06):1662-1668.

[13]高永梅,琚春华,邹江波.基于电信领域客户细分模型的个性化服务构建[J].数学的实践与认识,2015,45(05):44-54.

[14]邹轩. 基于缴费行为的电力客户细分及服务提升研究[D].宁波大学,2017.

[15]李伟,秦鹏,胡广勤,张毓福.基于商业大数据的客户分类方案[J].六盘水师范学院学报,2017,29(06):38-41.

[16]孙晓琳,姚波,陈瑜.基于客户资产离群数据分析的客户分类[J].统计与信息论坛,2018,33(10):114-120.

[17]陈明亮.基于全生命周期利润的客户细分方法[J].经济管理,2002(20):42-46.

[18]朱明英,邢豫,王海霞,王保中.基于客户价值的客户行为特征分类模型探讨[J].现代计算机(专业版),2017(01):7-10.

[19] Achim Walter,Hans Georg Gemunden，Thomans Ritter.Value Creation in buyer Seller Relationships[J].Industrial Marketing Managenment,2001(30):365-377.

[20] 李菊,王星,徐山.Prediction of customer classification based on rough set theory[J]. Procedia Engineering,2010(7):166-370.

[21]AdnanAmin，Feras Al-Obeidat，Babar Shah. Journal of Business Research[J]. Customer churn prediction in telecommunication industry using data certainty,2018 (94):290-301.

[22]Farid Shirazi,Mahbobeh Mohammadi.International Journal of Information Management[J].A big data analytics modle for customer churn prediction in the retiree segment,2018.

[23]王重仁,韩冬梅.基于社交网络分析和 XGBoost 算法的互联网客户流失预测研究[J].微型机与应用,2017,36(23):58-61.

[24]Kuanchin Chen,Ya-Han Hu,Yi-ChengHsieh.Information Systems and e-Business Management[J].Predicting customer churn from valuable B2B customers in the logistics industry: a case study, 2015.13 (3):475-494.

[25]程昊,樊重俊.神经网络在我国电商企业客户流失风险预测中的应用研究[J].经济研究导刊,2018(18):16-17.

[26]卢美琴,吴传威.商业银行贵宾客户流失预测研究[J].福建商学院学报,2018(02):31-36.

[27]卢光跃,王航龙,李创创,赵宇翔,李四维.基于改进的 K 近邻和支持向量机客户流失预测[J].西安邮电大学学报,2018,23(02):1-6.

[28]王建新. 符号型数据聚类算法的研究[D].山西大学,2016.

[29]王丽菊. 基于客户价值的航空旅客细分研究[D].北京邮电大学,2018.

[30]张良军、杨坦、徐圣兵等. MATLAB 数据分析与挖掘实战[M]北京：机械工业出版社，2016：97-99.

[31]Jiawei Han,Micheline Kamber.Data Mining Concepts and Techniques[M].北京：机械工业出版社.2012:247-254.

[32]孙亮，黄倩.实用机器学习[M].北京:人民邮电出版社 2017:286-291.

致谢

感谢培养教育我的贵州财经大学，学校学术氛围浓厚，学习环境舒适，我将终生难忘！三年求学生涯，在师长、亲友的支持下收获颇丰，在论文即将付梓之际，思绪万千。在此，谨向帮助、关心过我的老师、同学表示衷心的感谢。

首先，最感谢我的硕士生导师张文专教授。在张老师的帮助下，接触到很多大数据分析相关的内容，让我受益匪浅。在论文的完成过程中，张文专老师不仅从研究思路上对我进行启发，促使我进一步思考，而且在建立模型的过程中遇到问题时，张老师同我们一起研究、讨论，使论文工作可以顺利完成。在此，谨向张文专老师表示我最诚挚的谢意！

感谢余孝军、马赞甫、夏天等各位老师在论文开题、预答辩期间提出的宝贵意见。同时，也衷心感谢数统学院的各位领导与老师，各位老师深厚的学术造诣、严谨的治学风格、严肃的科学态度深深的令我折服。

感谢伴随我一起同窗苦读的同学们：陈思含、王泽珺同学等等，各位挚友给予我的鼓励和帮助，使我在学海徜徉中收获了一份倍感珍贵的友谊！

感谢我的父母，他们在精神和物质上的无私支持，坚定了我追求人生理想的信念。

攻读硕士学位期间科研成果

一. 论文：

[1]曹晓祎,申玉伟.FDI 对中国经济增长的影响——基于 VAR 模型[J].企业科技与发展,2018(09):12-13.

[2]申玉伟,曹晓祎.我国用电量的多元回归分析[J].电子技术与软件工程,2018(24):225.

二. 课题：

1.贵州省扶贫办研究课题：贵安新区精准扶贫工作成效第三方评估

2.贵州省统计局研究课题：贵州工业统计信息可视化

3.校级课题：贵州省财政收入预测

三.获奖：

1.“华为杯”第十四届中国研究生数学建模竞赛二等奖

2.第三届贵州省统计信息可视化大赛网页类作品一等奖、动态类作品二等奖、静态类作品二等奖