同济大学

TONGJI UNIVERSITY

《商务智能》课程论文（数据分析报告）

题目：猜您满意——航空服务旅客满意度相关预测分析

得分			
年级	2018 级	专业	信息管理与信息系统
姓名	蓝鑫	学号	1852804
姓名	林芳展	学号	1851122

猜您满意——航空服务旅客满意度相关预测分析

【摘要】本案例以某航空公司所收集的旅客满意度调查结果数据为研究对象，通过统计分析探究旅客年龄、舱位等级、在线订购便捷程度等 22 个相关因素对航空公司旅客满意度的影响，建立基于 CHAID 算法的决策树模型展示各个因素与旅客满意度之间的关联。结果表明，线上值机满意度、旅客类型等因素对旅客满意度有显著影响。
【关键词】旅客满意度，决策树算法，CHAID 算法

一、选题与背景介绍

随着全球旅游经济的不断发展，我国国民的出行量不断增多，飞机凭借其快捷、安全的优势，成为了人们外出旅游的常用交通工具之一。以 2019 年为例，中国民航完成旅客运输量 6.6 亿人次，同比增长 7.9%。民航旅客周转量在综合运输交通体系中占比达 32.8%，同比提升 1.5 个百分点。

但是，近年来高铁飞速发展，对航空公司的冲击越来越大。随着高铁的不断提速和完善，航空公司面临的挑战逐步加剧。如何在全球旅游业快速发展的时期，在全球各大航空公司的激烈竞争中，得到社会的认同，得到消费者的认可，从而获得更多的旅客，获得最大化经济效益成为航空企业面对的难题。因此，有些航空公司聚焦于提升旅客满意度这一可能手段，关于旅客满意度的研究便应运而生。

旅客满意度是指企业的所有产品对旅客一系列需求的实现程度，往往和顾客价值呈正相关关系，而企业的生存和发展离不开顾客价值。从经验和常识又可以猜测，服务质量、产品价格等因素可能直接或间接地影响旅客满意度。因此，通过调整旅客满意度的影响因素，可能可以实现旅客满意度的提升，进而提升顾客价值，为企业带来更多收益。

基于上述思考，我们认为相关研究具有较高的商用意义，因此确定如下选题：通过分析航空公司的旅客调研数据，对旅客满意度的影响因素进行探究，从而寻找能够影响旅客满意度的主要因素，对旅客满意度进行预测，为航空公司改善服务质量和提高服务水平提供具有较高性价比的行动建议。

二、数据来源与说明

本案例所使用的数据来自著名的机器学习网站 Kaggle，数据采集时间是 2020 年 2 月，共 129797 条数据（数据集本身已分为训练集和测试集），其中有 310 条为缺

失数据。考虑到数据量较大，将极少数数据删除对数据分析结果影响较小，因此对具有缺失值的数据做删除处理，剩余 103594 条训练集数据，25893 条测试集数据。数据集共包含 25 个变量，其中将旅客满意度作为因变量，将旅客因素（性别、旅行类型等）、航班因素（飞行距离、起飞延误时间等）和评价因素（座椅舒适度、腿部空间满意度等）共 22 个变量作为自变量，剩余的变量为序号变量与分类变量（训练集或测试集）。详细的变量说明表如表 2-1 所示。

变量类型		变量名	详细说明	取值范围	备注
因变量		旅客满意度	离散型变量	满意 /中立或不满意	满意取值为 1
自变量	旅客因素	旅客性别	离散型变量	男/女	男性取值为 1
		旅客类型	离散型变量	忠实旅客 /不忠实旅客	忠实取值为 1
		旅客年龄	单位：岁	7-85	只取整数
		出行类型	离散型变量	个人出行 /商务出行	出行目的 商务出行取 1
		舱位等级	离散型变量	商务舱/经济舱 /廉价舱	旅客的出行舱位分级
	航班因素	飞行距离	单位：英里	31-4183	未说明是否为累计值
		起飞延误时间	单位：分钟	0-1592	未说明是否为累计值
		降落延误时间	单位：分钟	0-1584	未说明是否为累计值
	旅客评价因素	机上 wifi 服务满意度	离散型变量	0-5	0 表示没有 1-5 表示满意度
		登机口位置	离散型变量	0-5	满意程度
		起飞/降落时间便捷度	离散型变量	0-5	便捷程度
		线上值机满意度	离散型变量	0-5	满意程度
		线上订票容易程度	离散型变量	0-5	容易程度
		食品和饮料	离散型变量	0-5	满意程度
		座椅舒适度	离散型变量	0-5	满意程度
		机上娱乐	离散型变量	0-5	满意程度
		机上服务	离散型变量	0-5	满意程度
		腿部空间满意度	离散型变量	0-5	满意程度
		行李处置满意度	离散型变量	1-5	满意程度
		检票服务满意度	离散型变量	0-5	满意程度
		飞行服务满意度	离散型变量	0-5	满意程度
		洁净程度	离散型变量	0-5	洁净满意度

表 2-1 数据变量说明表

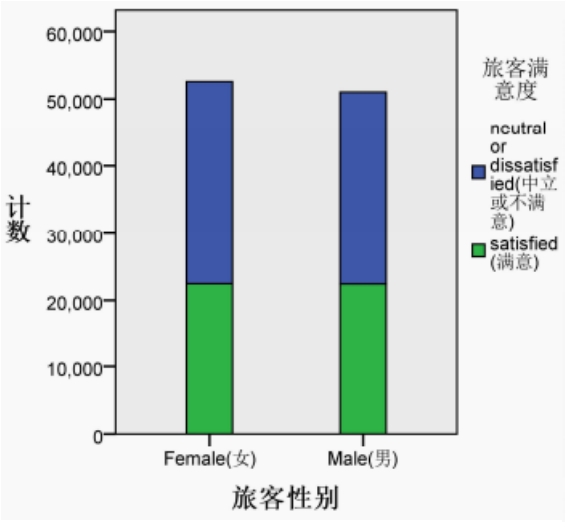
三、描述性分析

在对旅客满意度影响因素进行模型探究之前，首先对各变量进行描述性分析，以初步判断满意度的影响因素。为保证后续模型检验阶段的准确性，仅对训练集数据进

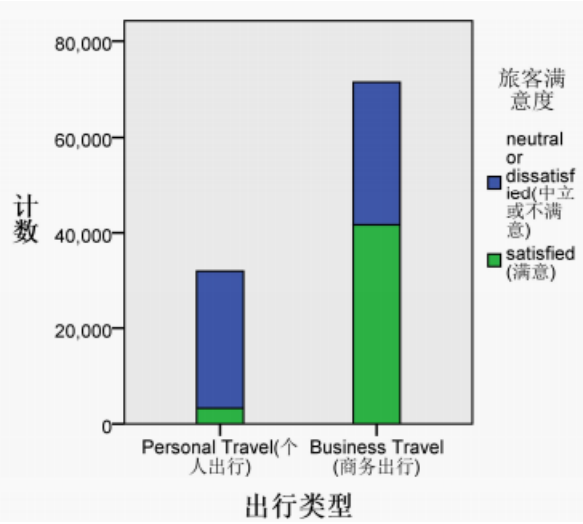
行分析。

（一）满意度与旅客因素的探究分析

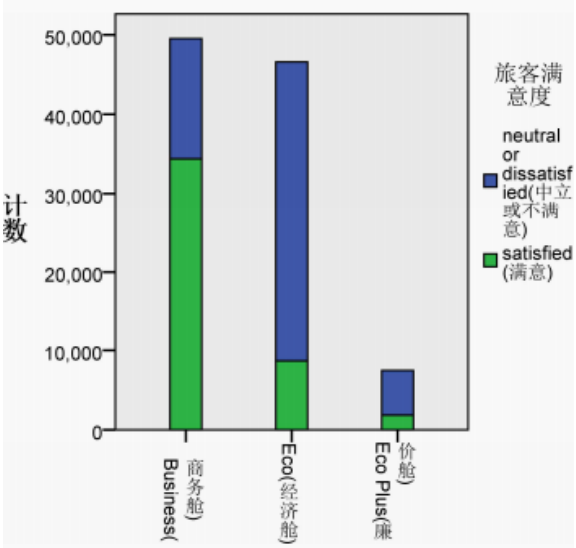
旅客因素包括旅客性别、出行类型、舱位等级、旅客类型和旅客年龄。以下为探究旅客因素与旅客满意度关系的一系列堆叠图和箱线图，其中纵坐标“计数”指对应项的旅客数量。



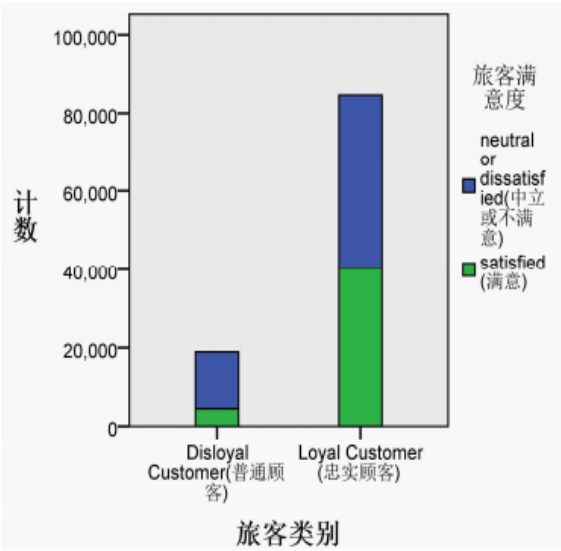
(a)旅客性别-旅客满意度堆叠图



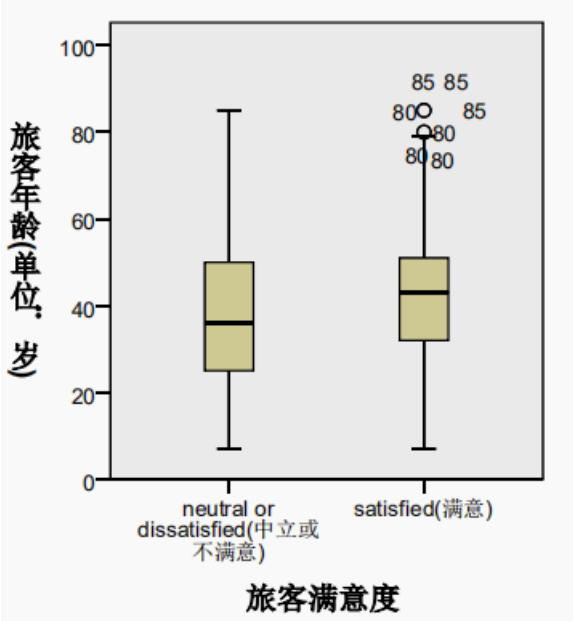
(b)出行类型-旅客满意度堆叠图



(c)舱位等级-旅客满意度堆叠图



(d)旅客类型-旅客满意度堆叠图



(e)旅客年龄-旅客满意度箱线图

图 3-1 旅客因素与旅客满意度关系系列图

通过图3-1可以初步得到一些结论。(a)旅客性别-旅客满意度堆叠图显示，男女旅客满意度比例差异不大，同样在(e)旅客年龄-旅客满意度箱线图中，不同旅客满意度对应的旅客群体的年龄差别不大，说明性别或旅客年龄可能不是旅客满意度重要的影响因素。

另外，(d)旅客类型-旅客满意度堆叠图显示，不太忠诚的旅客对航空服务不满的比例远大于忠诚旅客的。当然，不满本身可能也是导致旅客不忠诚的一个原因，进一步深究需要得知航空公司是以何种手段判定旅客忠诚与否的，但提供该数据集的航空公司没有给出明确说明。

从(b)出行类型-旅客满意度堆叠图中可以发现，个人类型的旅客的不满比例远大于商务类类型的，且该类型基数不小。在(c)舱位类别-旅客满意度堆叠图中，同样可见类似的显著比例区别，且不满比例较高的经济舱类型对应的旅客基数也很大。这意味着改善个人类型航班或乘坐经济舱的旅客的体验可能是该公司提高旅客满意度的努力方向。

（二）满意度与航班因素的探究分析

航班因素包括飞行距离、起飞延误时间和降落延误时间。以下为探究航班因素与旅客满意度关系的堆叠图和描述性分析表格。

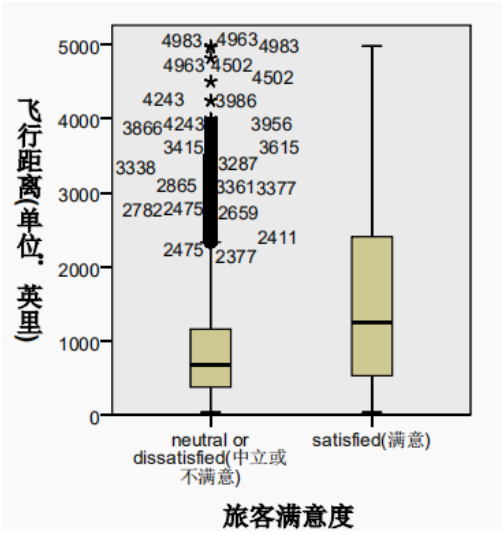


图3-2 飞行距离-旅客满意度箱线图

		起飞延误时间	降落延误时间
个案数	有效	44897	44897
	缺失	0	0
平均值		12.55	12.63
平均值标准误差		.167	.170
中位数		.00	.00
众数		0	0
标准 偏差		35.316	35.962
方差		1247.200	1293.266
偏度		7.404	7.176
偏度标准误差		.012	.012
峰度		114.429	104.276
峰度标准误差		.023	.023
范围		1305	1280
最小值		0	0
最大值		1305	1280
总和		563294	567085
百分位数	25	.00	.00
	50	.00	.00
	75	9.00	8.00

a. 训练集或测试集 = train(训练集)

(a)满意旅客起飞/降落延误时间对比(单位：分钟)

		起飞延误时间	降落延误时间
个案数	有效	58697	58697
	缺失	0	0
平均值		16.43	17.13
平均值标准误差		.165	.167
中位数		.00	.00
众数		0	0
标准 偏差		40.046	40.560
方差		1603.694	1645.134
偏度		6.390	6.256
偏度标准误差		.010	.010
峰度		93.769	88.682
峰度标准误差		.020	.020
范围		1592	1584
最小值		0	0
最大值		1592	1584
总和		964504	1005335
百分位数	25	.00	.00
	50	.00	.00
	75	15.00	17.00

a. 训练集或测试集 = train(训练集)

(b)不满意/中立旅客起飞/降落延误时间对比(单位：分钟)

表3-1 起飞/降落延误时间与旅客满意度描述性分析表

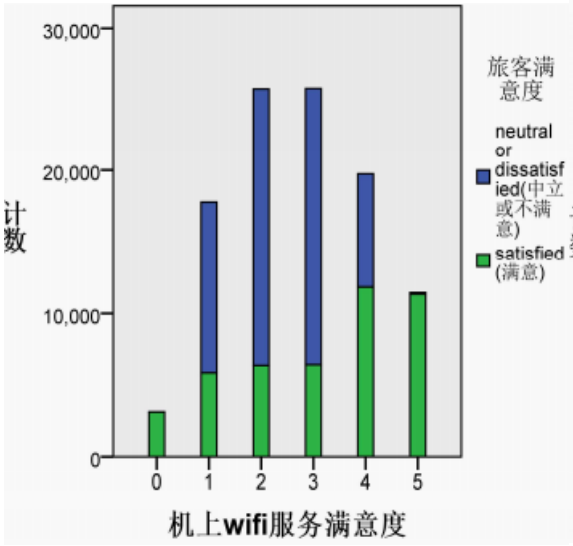
由图3-2可以看出，整体来讲，不满意的旅客飞行距离较短，但从常识推断，这很可能是因为满意的旅客会多次选择该航空公司的航班，而不一定是因为飞行距离短会增加旅客对航空服务不满意的概率。但是，数据集未说明飞行距离是“单次”还是“累计”的，上述猜想是在飞行距离是“累计”的前提下作出的。

从表 3-1 的对比来看，满意的旅客的起飞/降落延误时间总体平均值低于不满意

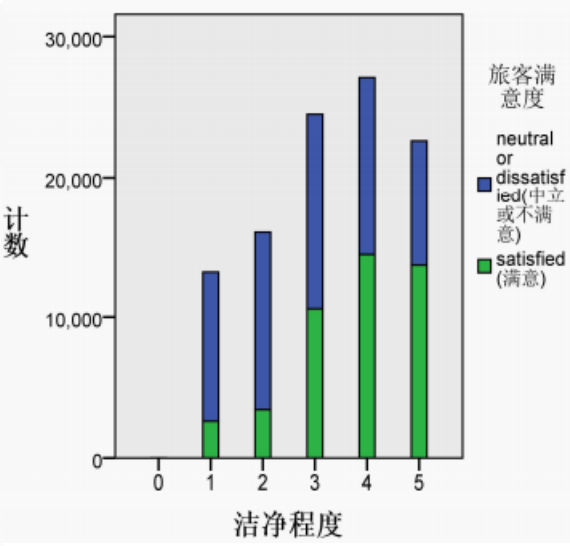
的旅客，可能说明旅客经历的两类延误时间越长，旅客满意度可能越低，这个猜测较符合常识。但值得注意的是，数据集对应的满意/不满意或中立的旅客中均有半数未经历过任何类型的延误，这意味着旅客经历延误本身可能已经是一个较小概率的事件，因此旅客经历的延误时间对旅客满意度的预测意义可能较小。

（三）满意度与旅客评价因素的探究分析

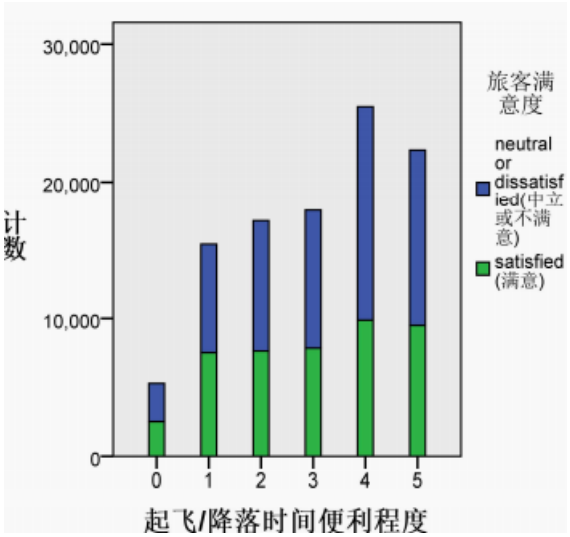
以下为探究旅客评价航空公司服务的某方面的满意度/便捷度/容易度与旅客满意度关系的一系列堆叠图，其中纵坐标“计数”指对应项的旅客数量，横坐标为旅客给出的各个评分评级。



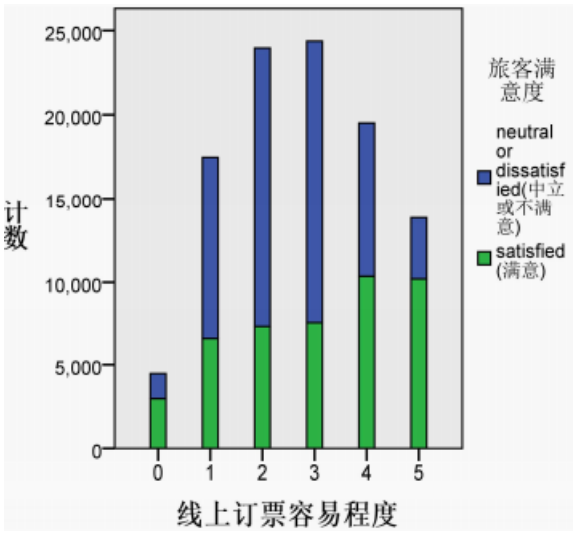
(a)机上wifi服务满意度-旅客满意度堆叠图



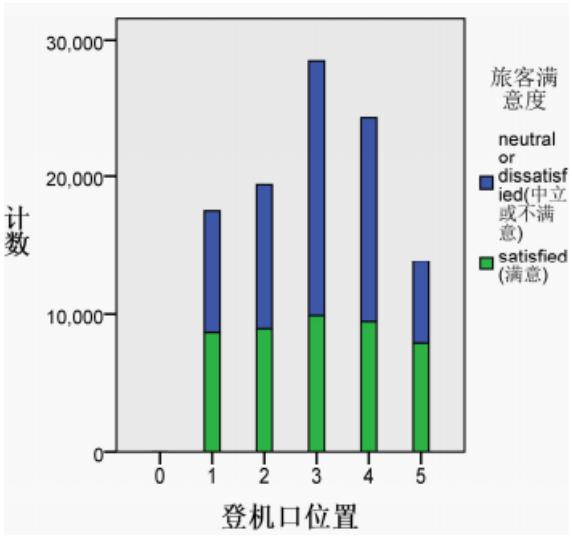
(b)洁净程度满意度-旅客满意度堆叠图



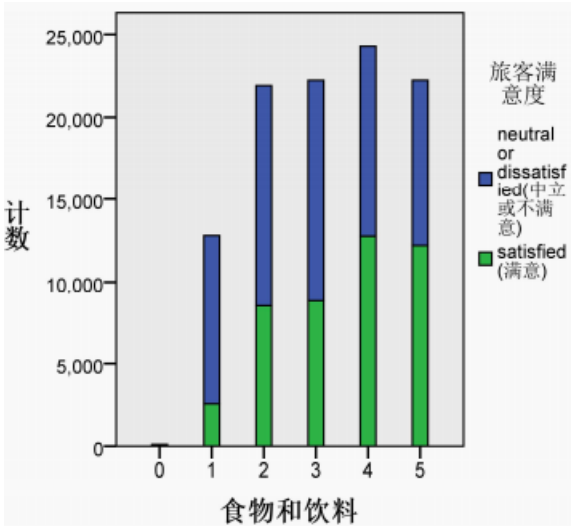
(c)起飞/降落时间便利程度-旅客满意度堆叠图



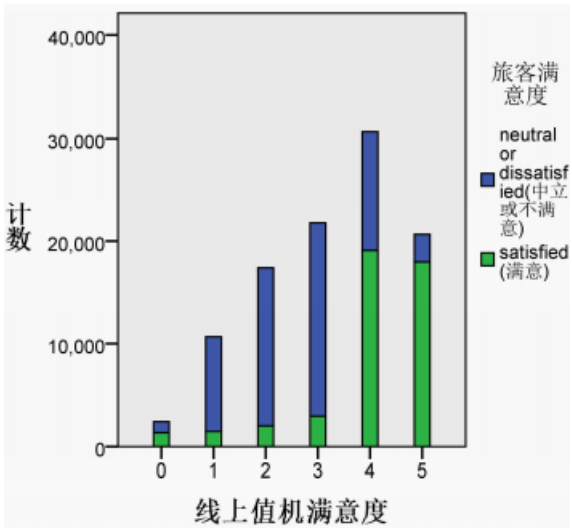
(d)线上订票容易程度-旅客满意度堆叠图



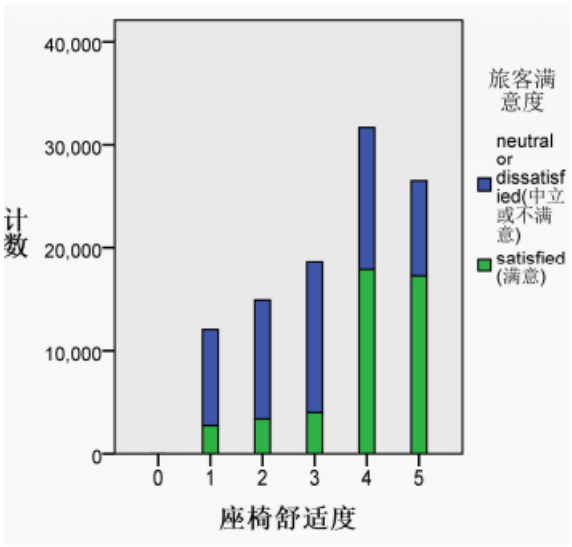
(e)登机口位置满意度-旅客满意度堆叠图



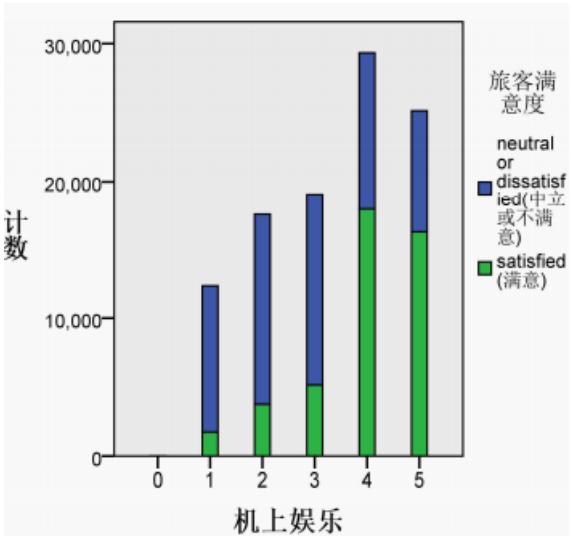
(f)食物和饮料满意度-旅客满意度堆叠图



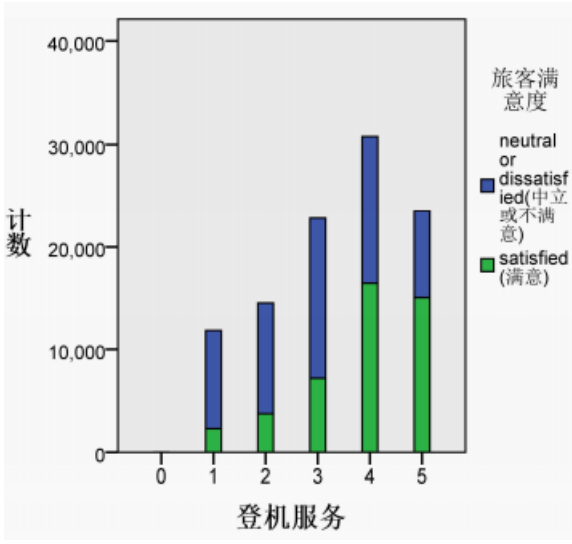
(g)线上值机满意度-旅客满意度堆叠图



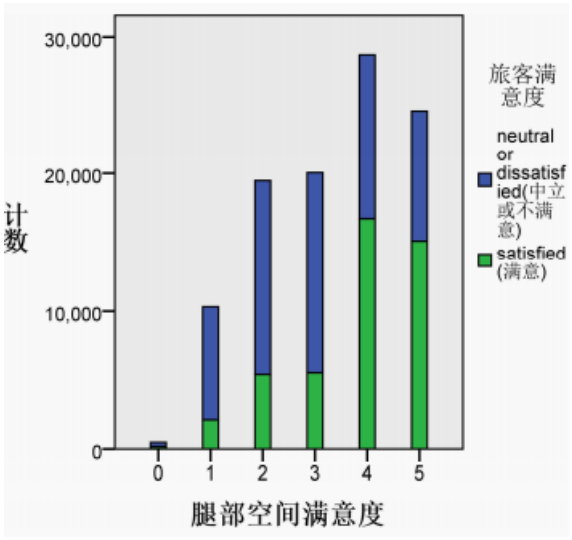
(h)座椅舒适度-旅客满意度堆叠图



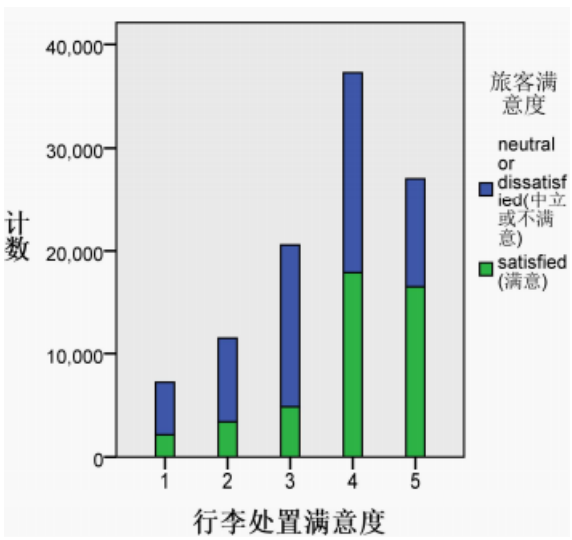
(i)机上娱乐满意度-旅客满意度堆叠图



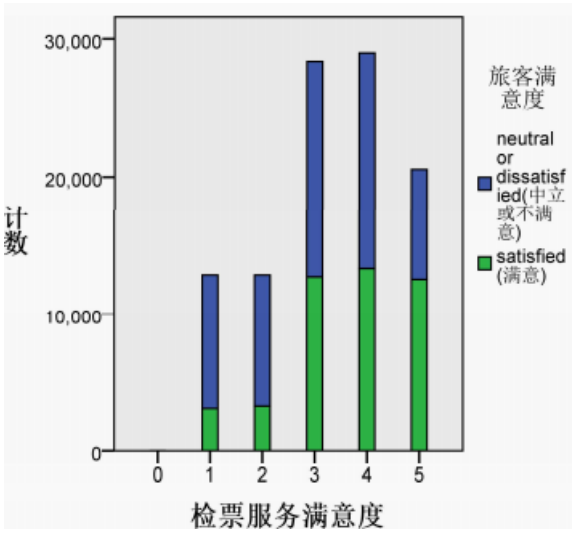
(j)登机服务满意度-旅客满意度堆叠图



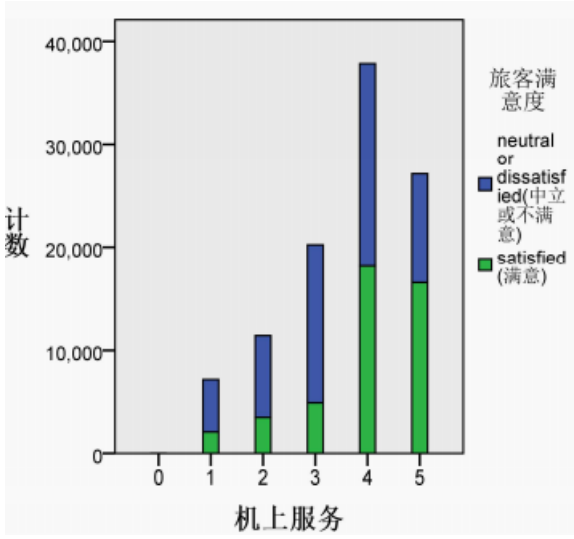
(k)腿部空间满意度-旅客满意度堆叠图



(l)行李处置满意度-旅客满意度堆叠图



(m)检票服务满意度-旅客满意度堆叠图



(n)机上服务满意度-旅客满意度堆叠图

图 3-3 旅客评价因素与旅客满意度关系系列图

通过对图像的观察，可以简要地概括这些顾客评价因素和总体满意度的关系的规律：对任意一个方面的旅客评价满意度，评级越高，旅客满意的比例往往也越高。这是相对符合常识的。

当然，在这些顾客评价因素中，存在一些相对反常的情况。机上 wifi 服务满意度评分为 0 时，旅客满意比例相当高（数据集说明了该评分意味着对应飞机上不提供 wifi 服务）。线上订票服务满意度评分为 0 时，同样出现了高于其他评分分级的旅客满意比例（但数据集未说明线上订票服务满意度评分为 0 的意义）。另外，旅客对登机口位置的评分为 3 时，旅客满意比例比其他评分分级的要低。这些情况可能具有特殊意义。

（四）相关系数矩阵

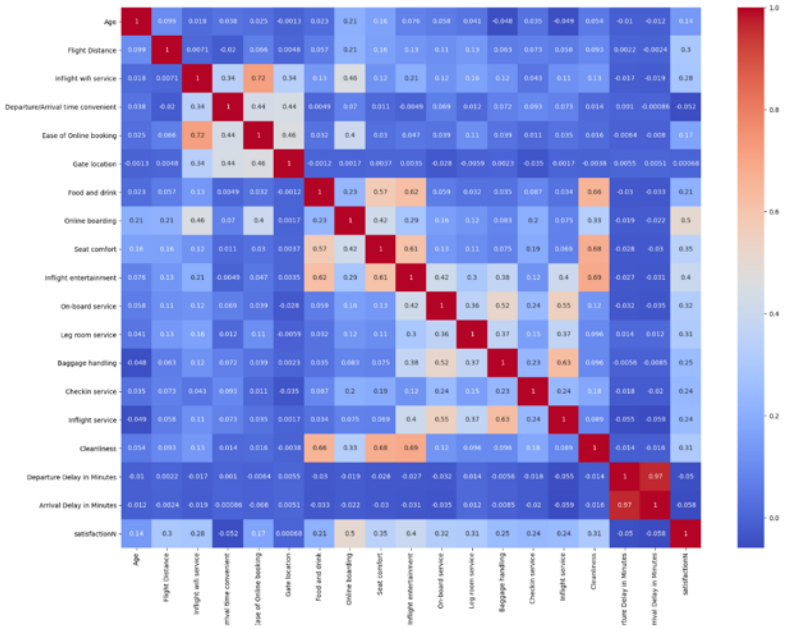


图 3-4 因变量及自变量相关系数矩阵图（高清图片请见附录文件）

从 Pearson 相关系数矩阵可以看出，与总体满意度相关性较高(r>=0.3)的变量有飞行距离、线上值机满意度、座椅舒适度、腿部空间满意度、洁净满意度等。与常识略有相悖的是，各类延误时间与旅客满意度相关性均很低，这可能是因为延误本身就是一个比较稀有的事件，大部分旅客未遇到过延误情况，因此相关性不强。值得注意的是，两类延误时间的相关性相当高。

然而，在进行后续建模分析时，仍应考虑不同自变量之间的相互影响关系，因此不能在建模时简单舍弃与总体满意度相关性较低的单个变量。

四、建模与分析

为了更深入地分析各因素对旅客满意度的影响，本案例将建立旅客满意度关于旅客因素、航班因素、旅客评价因素的决策树模型，并试图借助该模型进行一系列的应用。选择决策树算法可以规避多重共线性可能带来的问题，提供解释性高的模型。

（一）基于 CHAID 算法的决策树模型

我们通过 SPSS 的分析模块，建立了基于 CHAID 算法的决策树模型，通过训练集对模型进行训练，并使用测试集的数据验证模型的准确性(具体语法请见附录文件)。决策树模型摘要结果如表 4-1 所示，决策树训练结果如图 4-1 所示。

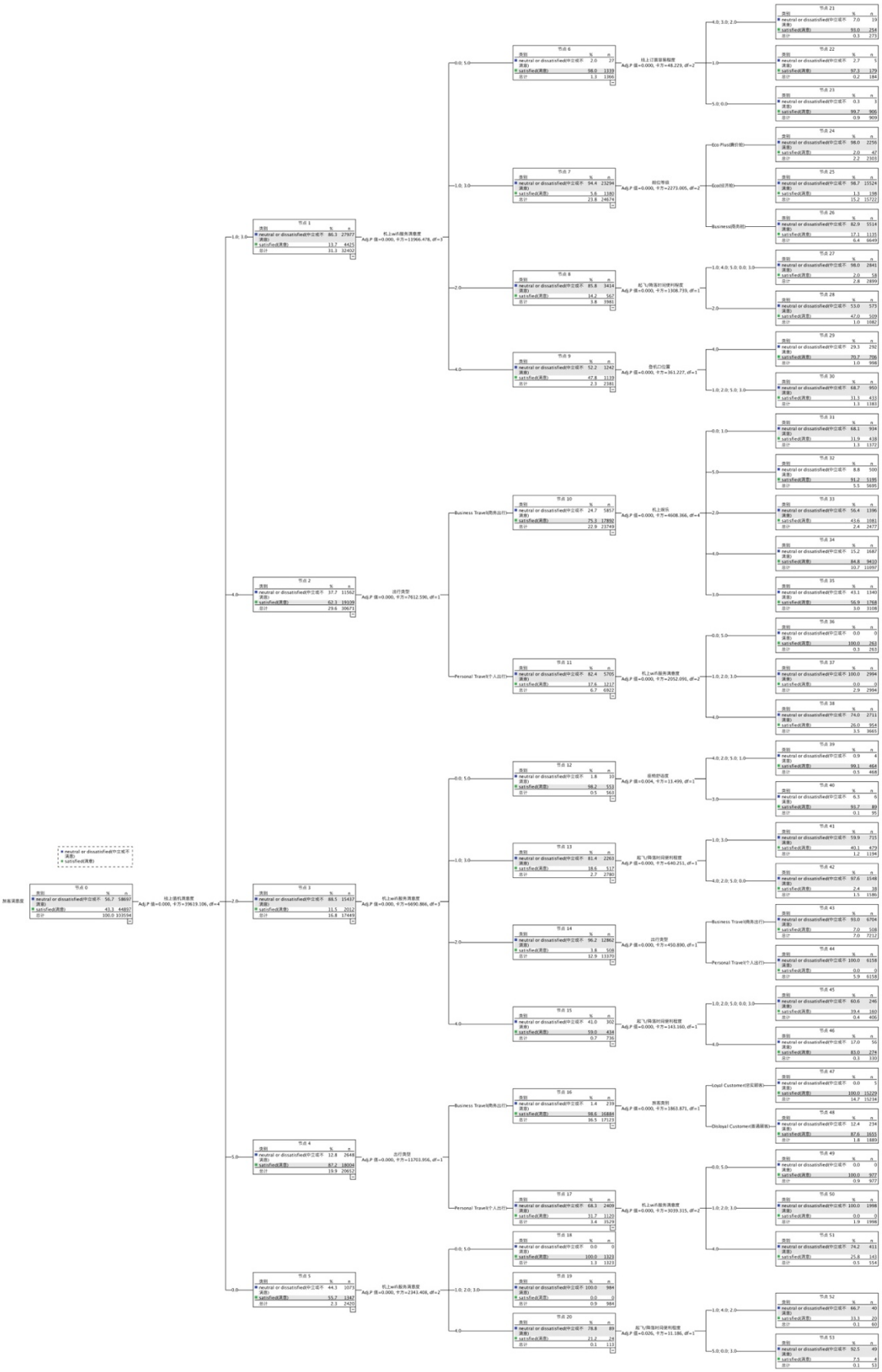


图 4-1 决策树模型（高清图片请见附录文件）

规格	生长法	CHAID
	因变量	旅客满意度
	自变量	旅客性别，旅客年龄，旅客类别，出行类型，舱位等级，飞行距离，机上 wifi 服务满意度，起飞/降落时间便利程度，线上订票容易程度，登机口位置，食物和饮料，线上值机满意度，座椅舒适度，机上娱乐，登机服务，腿部空间满意度，行李处置满意度，检票服务满意度，机上服务，洁净程度，起飞延误时间，降落延误时间
	验证	拆分样本
	最大树深度	3
	父节点中的最小个案	100
	子节点中的最小个案	50
结果	已包括自变量	线上值机满意度，机上 wifi 服务满意度，线上订票容易程度，舱位等级，起飞/降落时间便利程度，登机口位置，出行类型，机上娱乐，座椅舒适度，旅客类别
	节点数	54
	终端节点数	35
	Depth	3

表 4-1 决策树模型摘要

在表 4-1 中可以看出，SPSS 的算法规则剔除了一些分类意义较小的自变量（例如：节点对应个案数需大于 50 个，否则不进行分类等规则）。对图 4-1 的模型结果进行分析后，可以得出以下结论：

1.旅客因素：根据表 4-1 所得的结论中，与旅客满意度关联较大的因素有旅客类别、舱位等级和出行类型，与旅客的性别、年龄关系不大。

2.旅客评价因素：旅客满意度与旅客评价因素之间的关系大致呈现出，各项评价指标越高（越接近于 5），旅客越能感到满意的规律。例如，在一定条件的约束下，机上娱乐服务评价越高，旅客越能感受到满意。

这些结论与之前的猜想基本符合，并且模型中的决策结果 p 值均小于 0.05，可以认为模型的准确度较高，拟合程度可以接受。

下面利用测试集进行模型检验，如图 4-2 与表 4-2 所示。

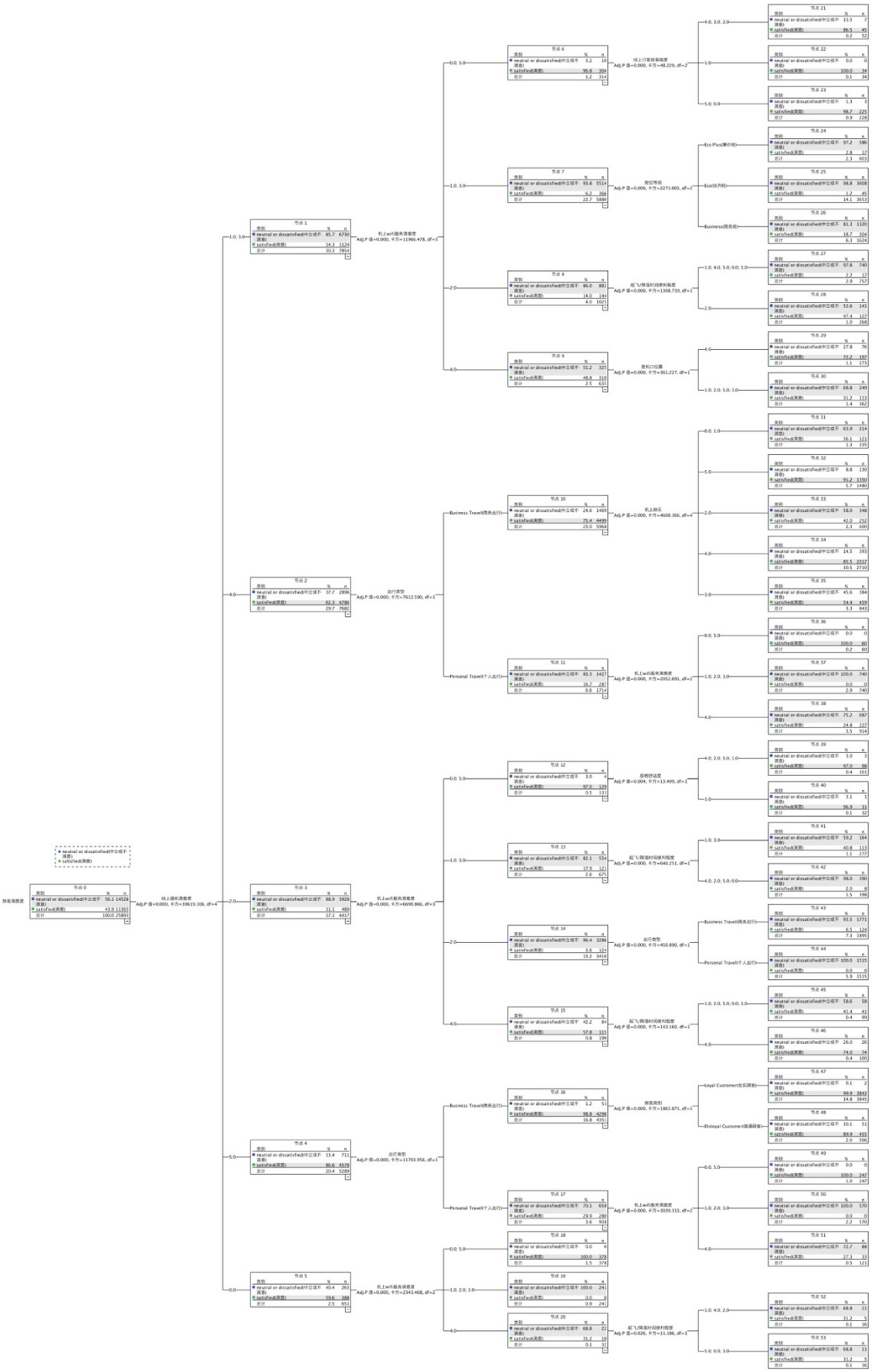


图 4-2 决策树测试集检验结果（高清图片请见附录文件）

样本	观测值	观测值		
		中立或不满意	满意	正确百分比
训练	中立或不满意	54546	4151	92.9%
	满意	6205	38692	86.2%
	总体百分比	58.6%	41.4%	90.0%
检验	中立或不满意	13452	1076	92.6%
	满意	1552	9813	86.3%
	总体百分比	57.9%	42.1%	89.9%

表 4-2 决策树模型训练与检验结果

从图 4-2 决策树的测试集检验结果中可以看出，没有明显的偏离模型的现象，并且每个节点的卡方检验 p 值几乎均小于 0.01，只有一个节点为 0.026。另外，表 4-2 说明测试集总体的预测正确率为 89.9%，属于较高水平，因此我们认为使用该模型刻画旅客满意度与各因素的关系是比较合理的。

（二）模型预测应用

我们使用基于 CHAID 算法的决策树模型进行分析，对本案例所采用的数据具有较强的解释性，可利用决策树基于某个旅客在各变量上的数值给出对该旅客满意度的准确度较高的预测。

基于模型结果，我们向航空公司给出以下行动建议：对于各个航空公司而言，除了降低机票价格、增设航线这些常见做法外，想要牢牢把握住旅客，应当在不同的角度下多花功夫，提升旅客满意度。根据本模型所得到的结果，航空公司可以通过对不同的服务或设施进行改善，如开发更容易上手的线上值机软件、提供优质的 wifi 服务等提升旅客满意度，并且在后续的改善进程中，不断收集新的数据与旅客的诉求，进一步开发和完善决策树模型。

五、结论与建议

本案例对某航空公司所收集的旅客满意度调查结果数据进行了统计分析，得到如下的结论。

影响旅客满意度的主要因素有：

1.旅客因素：出行类型、旅客类型、舱位等级。

2.旅客评价因素：线上值机满意度，机上 wifi 服务满意度，线上订票容易程度，起飞/降落时间便利程度，登机口位置，机上娱乐，座椅舒适度。

可根据上述影响因素的取值基于决策树模型对旅客满意度进行预测。

有很多影响旅客满意度的因素在本报告中因SPSS决策树模型处理规则自动排除，若采用其他算法可纳入考虑。另外，在未来的研究与改进中可以考虑获取更多类型的影响因素数据以供分析，如机票价格、旅客邻座是否有婴儿等。在实际应用中，如果要将该模型推广至更多的航空公司，可能还要进一步考虑不同航空公司所采用的机型等具有较高独特性的特殊因素的影响。

由于本案例的数据来源经过脱敏处理，因此部分内容的具体来源和含义无法得知，对模型的解释性造成了一定影响，也难以基于成本进行分析以给出更具体的行动方案。对航空公司而言，可以收集更为精确、来源更清晰的数据，再利用本案例的方法加以建模，即可能获得拟合程度更好、解释性更高的结果。

附录

[1]train.csv 原始训练集数据

[2]test.csv 原始测试集数据

[3]data.sav/data.csv 经处理后的数据

[4]cor.png 相关系数矩阵高清图片

[5]训练模型.eps/训练模型.pdf 决策树训练模型高清图片

[6]测试模型.eps/测试模型.pdf 决策树测试模型高清图片

[7]决策树语法.sps 生成决策树对应的SPSS语法文件

[8]决策树输出.spv 决策树模型对应的一系列SPSS输出结果

参考文献

[1]潘海全. 南航广西分公司客舱服务顾客满意度研究[D].桂林理工大学,2020.

[2]马芳芳,陈秋萍.双因素理论视角下的航空服务质量旅客满意度研究:以厦门航空为例[J].西南交通大学学报(社会科学版),2020,21(06):101-110.