《属性数据分析》课程之数据分析报告

某航空公司乘客对旅途的满意度数据分析

姓名：梅新宇

学号：19307100079

日期：2022 年 12 月 31 日

摘要

在航空业竞争日益激烈、中国出入境即将恢复的当下，服务质量和口碑对航空公司来说越来越重要。因此，本文收集并分析了某航空公司乘客评价的数据。首先，采用探索性分析研究数据集各变量的分布和关系，其中，根据到达延误时间和出发延误时间这两个变量的线性关系，填充了到达延误时间的空值；其次，初步检查自变量与因变量的相关性来试图降维，对于连续型变量，使用 t 检验分析不同评价下的差异；然后，采用卡方检验研究属性变量与因变量的独立性；之后建立 logistic 回归模型定量研究影响乘客对旅途满意与否的因素；最后，应用 KNN 最近邻算法和随机森林算法进行拓展分析。

主要结论及研究成果包括以下几个方面：（1）长距离的航班满意度更高；(2)乘客类型、年龄、出行原因、出发延误时间、舱位、出发到达时间是否便利、网上订票是否容易、线上值机、腿部服务、行李处理、登机服务对乘客满意度的影响较为显著；（3）忠实乘客、年轻乘客、商务出行都会提高乘客对行程满意的概率；（4）是否商务出行和舱位是重要性最大的两个变量；(5)建立三个独立预测模型，训练精度可分别达到 89.32%、71.61%和 96.30%。基于以上研究成果，我们从设定目标客户、改善航班安排、提高线上服务等角度出发，为航空公司的经营提供了合理化的建议。

关键词：航空公司；服务；乘客特征

目录

第一章 问题描述

1.1 数据集来源

本文选取的数据集记载了美国某航空公司的乘客对旅途满意度的调查数据，共包含 129880 条观测值，有乘客的年龄、舱位、对细分服务项目满意程度、对旅途总体满意或不满意等 25 个变量。数据集来源于 Kaggle 网站：https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction，上传者为 TJ KLEIN。在网站上，数据集已经按照 4：1 划分为了训练集和测试集。

1.2 研究意义

现在，航空业非常成熟，在飞行技术上已经走到平台期，航空公司要靠飞行技术的突破来换取竞争力需要不少的时间和运气，是比较困难的。因此，为了提高在行业内的竞争力，航空公司都锁定在 C 端，希望提高航程服务的质量和口碑。

此外，近日中国卫健委发布通知，2023 年 1 月 8 日起不再对入境中国的旅客实行防疫隔离，且中国公民的海外旅游也将逐步重新启动。这对于中国的航空业来说，是一个积极的复苏信号。在经历了三年的低迷之后，各大航空公司也一定希望以更好的面貌迎接世界各地的乘客。

本文通过分析航空公司乘客满意度数据集，得到影响乘客是否对旅途满意的重要因素，能够为航空公司改善服务质量提供有价值的建议。

第二章 探索性分析

2.1 数据集概况

在第二第三章中的研究都针对整个数据集进行。该数据集共有 129880 条观测值，24 个变量，其中 23 个为自变量。其中，Inflight wifi service（航行中 wifi 服务）等 14 个变量是评价乘客主观满意度的重要变量，由乘客按 0-5 打分，分数越高，代表越满意。虽然这些变量的数值大小能对应乘客对该项服务满意的程度，但各满意程度之间的差距比较模糊且个性化，并不像数值一样等差，因此

本研究仍然将这 14 个变量视为分类变量。

另外，有 393 条观测值在 Arrival Delay in Minutes（到达延误时间）这个变量上是空值；其余变量均没有空值。对比 129880 的总观测数，空值并不算多，因此之后先在描述性统计中初步推测该变量与其他变量的关系，希望能够找到变量间的可靠数值关系用来填充空值，否则就直接用该变量非空值的中位数来填充。

表格 1 数据集中所有变量介绍

变量	解释
id	分类变量，乘客的 id。
Gender	分类变量，乘客的性别。取值为 Male/Female，后期为方便建模，修改为 0/1
Customer Type	分类变量，是否忠诚客户。取值为 Loyal/disloyal Customer，后期为方便建模，修改为 0/1
Age	数值变量，乘客的年龄
Type of Travel	分类变量，出行原因。取值为 Personal/Business Travel，后期为方便建模，修改为 0/1
Class	分类变量，舱位。取值为 Eco/Eco Plus/Business，后期为方便建模，修改为 3 个 0/1 变量
Flight Distance	数值变量，旅途距离
Inflight wifi service	分类变量，对飞行中无线网服务的满意度。取值为 0/1/2/3/4/5
Departure/Arrival time convenient	分类变量，对出发/到达时间的便利程度的满意度。取值为 0/1/2/3/4/5
Ease of Online booking	分类变量，对网上订票简易性的满意度。取值为 0/1/2/3/4/5
Gate location	分类变量，对登机口位置的满意度。取值为 0/1/2/3/4/5
Food and drink	分类变量，对飞行中食物和饮料的满意度。取值为 0/1/2/3/4/5
Online boarding	分类变量，对网上值机的满意度。取值为 0/1/2/3/4/5
Seat comfort	分类变量，对飞行中座椅舒适度的满意度。取值为 0/1/2/3/4/5

Inflight entertainment	分类变量，对飞行中娱乐活动的满意度。取值为 0/1/2/3/4/5
On-board service	分类变量，对登机服务的满意度。取值为 0/1/2/3/4/5
Leg room service	分类变量，对飞行中腿部服务的满意度。取值为 0/1/2/3/4/5
Baggage handling	分类变量，对行李处理的满意度。取值为 0/1/2/3/4/5
Checkin service	分类变量，对登机手续服务的满意度。取值为 0/1/2/3/4/5
Inflight service	分类变量，对飞行中服务的满意度。取值为 0/1/2/3/4/5
Cleanliness	分类变量，对飞机整洁度的满意度。取值为 0/1/2/3/4/5
Departure Delay in Minutes	数值变量，出发延误的时间。单位为分钟
Arrival Delay in Minutes	数值变量，到达延误的时间。单位为分钟
satisfaction	（因变量）分类变量，乘客对本次旅途是否满意。取值为 neutral or dissatisfied/satisfied，后期为方便建模改为 0/1

2.2 属性变量描述性统计

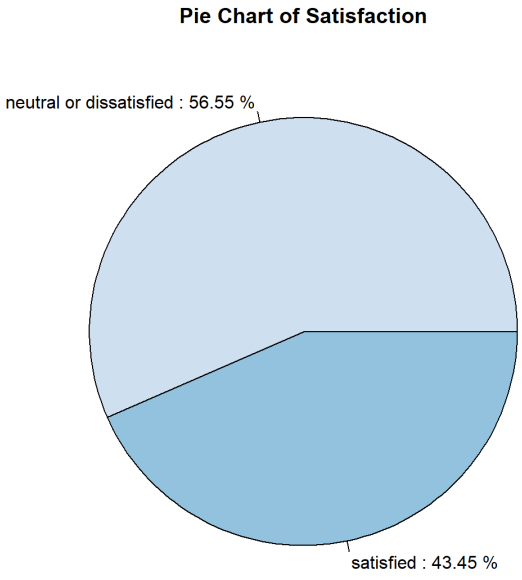

Figure 1 **因变量取值饼图**

由图 1 可知，因变量分布较为均衡，满意/中立或不满意的数据均接近 50%，样本合理。

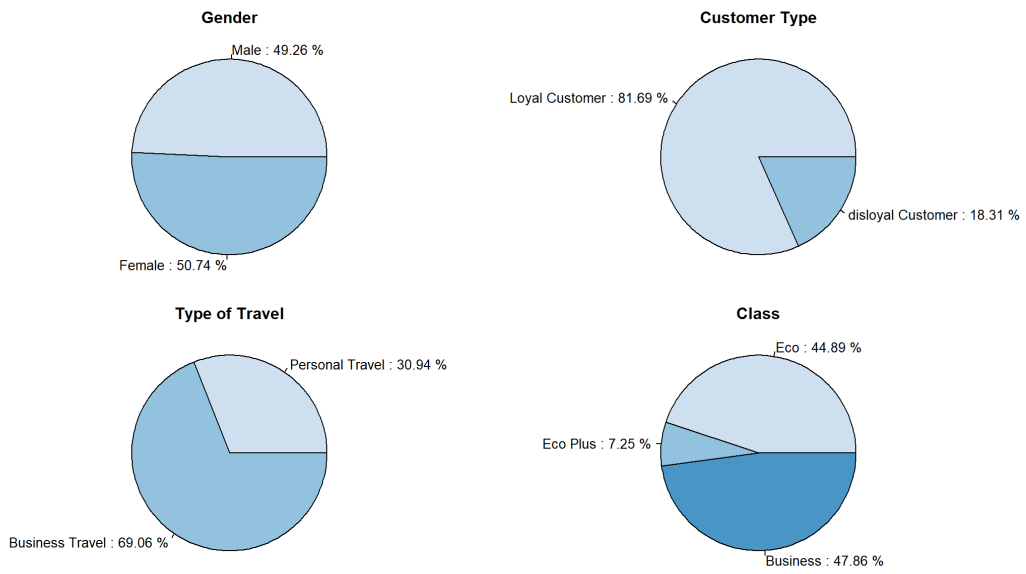

Figure 2 **乘客信息饼图**

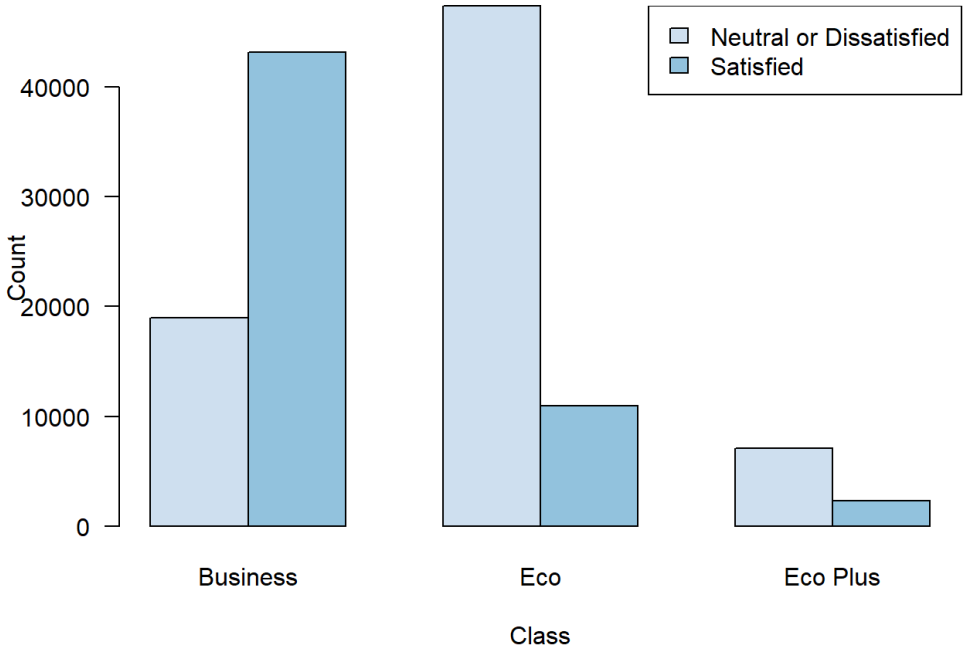

Figure 3 **舱位*是否满意分组条形图**

由图 2 可知，数据集在性别上的分布均衡。多数乘客为忠实顾客，即多次搭乘该航空公司的航班。多数观测值为商务出行的乘客，这与人们本身的需求有关，出差等商务出行的频率的确会比旅游、探亲等私人出行的频率高很多。另外，多数旅客乘坐的是商务舱或经济舱，且两者比例较为均衡，这使得服务评价数据的有效性较高，因为商务舱和经济舱接受的服务质量本身有一些差别（图 3 验证了这一说法）。

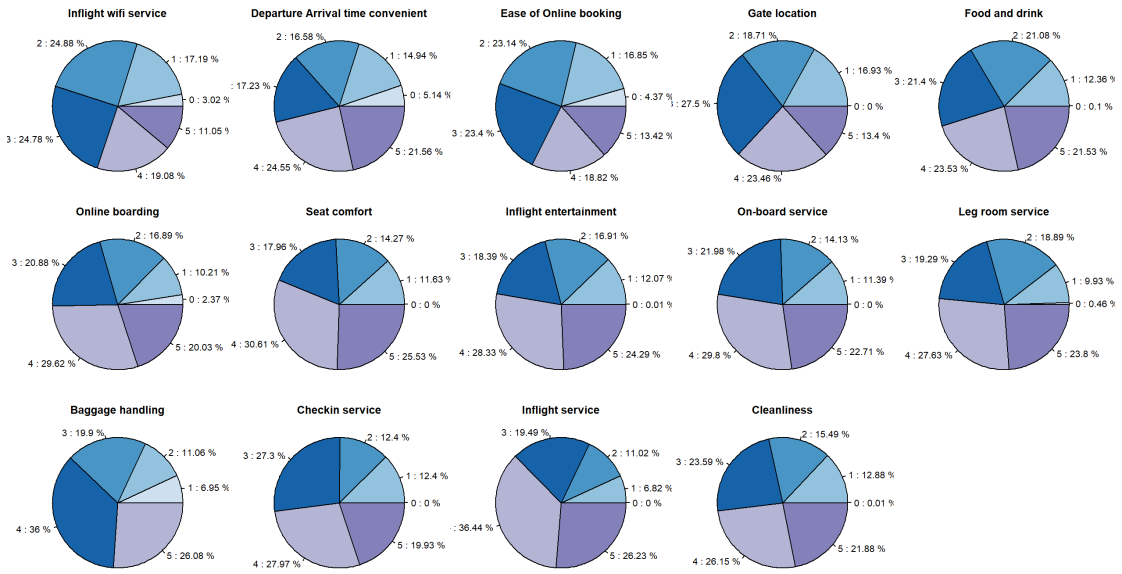

Figure 4 **各服务满意度饼图**

由图 4 可知，多数服务的分数集中在 2-5 分，如果将 0-5 分分为三档 0-1，2-3，4-5，分别解读为不满意、普通、满意，则针对特定服务评价时，乘客不太倾向于选择不满意。不过，前面提到样本中 satisfaction 这一字段中，选择满意/中立或不满意的乘客人数相当，从生活经验来看，这可能是因为哪怕仅有一项服务不够满意，乘客的心情就会被影响，从而认为整个旅途不令人满意。

2.3 数值变量描述性统计

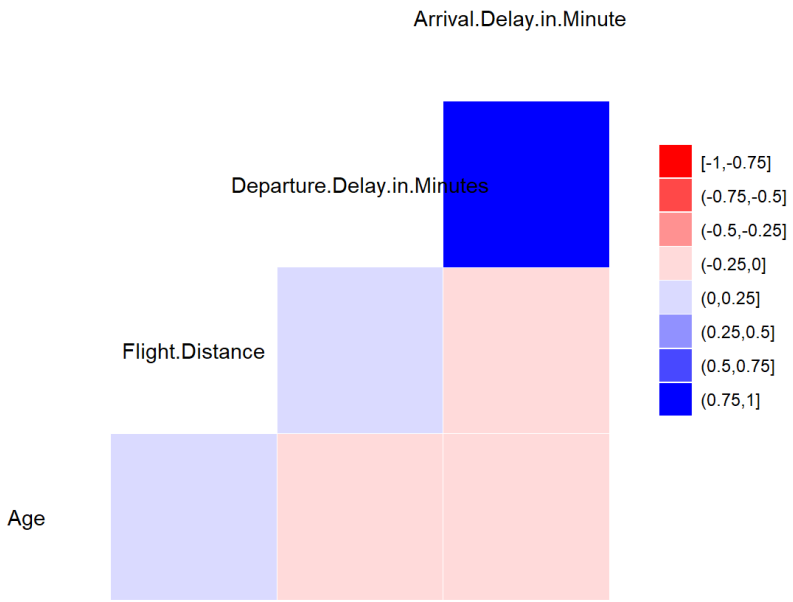

Figure 5 **数值变量相关系数热力图**

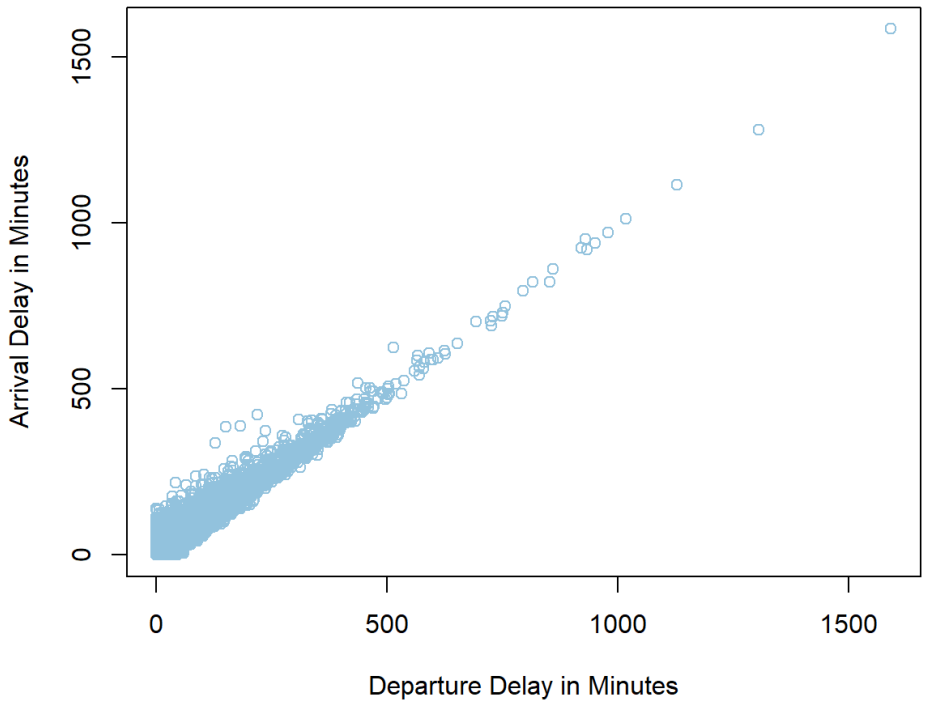

Figure 6 **到达延误时间和出发延误时间散点图**

首先，通过图 5、图 6 可知，在数值变量中，到达延误时间与出发延误时间之间有非常强的线性相关性，而其他变量之间则没有。根据经验来说，这是非常合理的结果，因为出发时延误时间越长，到达时间也会与最初预估的有越大偏差。有了这一关系，我们可以根据数量关系来填充到达延误时间这个变量的空值，也可以在建模中只选两者中的一个，来达到降维的效果。

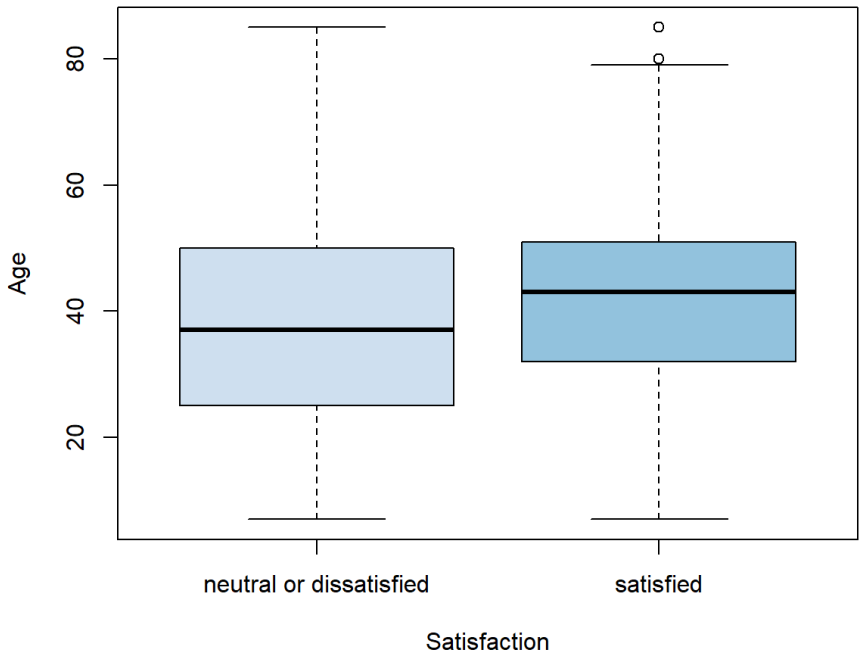

Figure 7 **年龄关于是否满意箱线图**

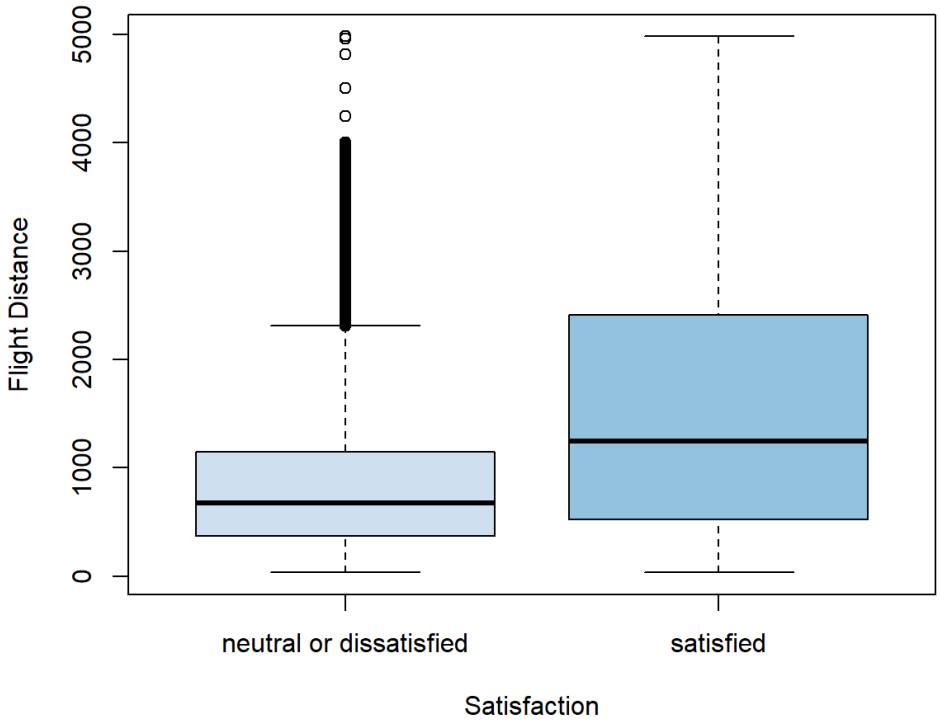

Figure 8 **航程关于是否满意箱线图**

由图 7 和图 8 可以看到对航程满意的乘客年龄比中立或不满意的分布更集中，偏大。从飞行距离来看，对旅途满意的乘客航程分布更广，且显然比中立或不满意的距离长。换句话说，旅途越长，越容易满意。

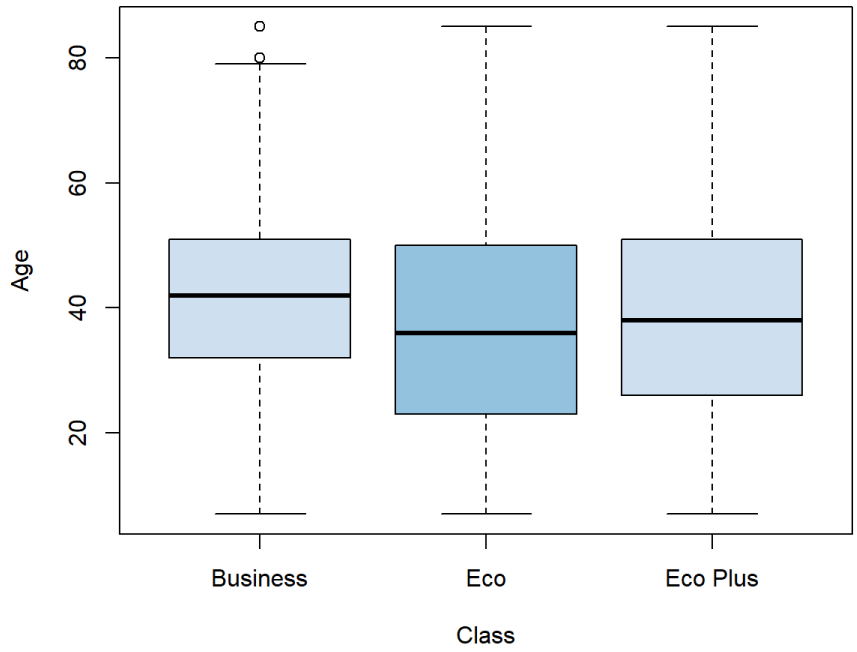

Figure 9 **年龄关于舱位箱线图**

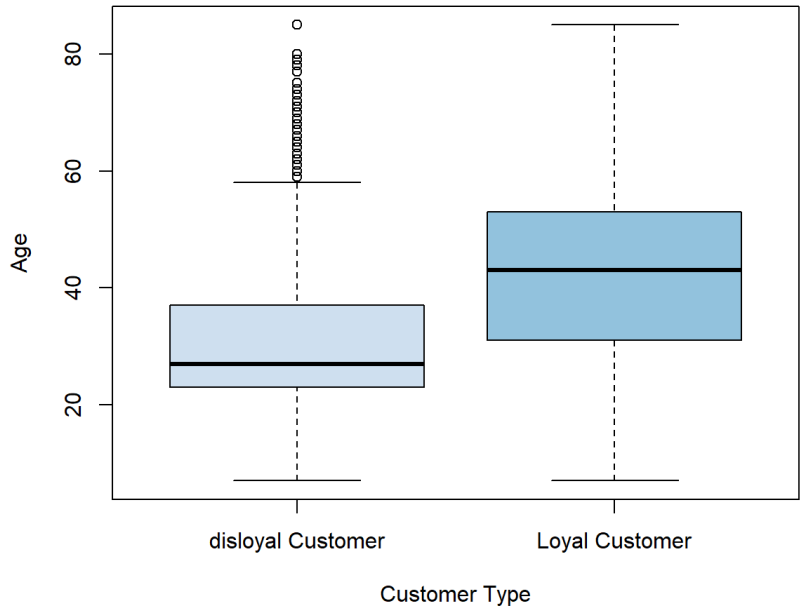

Figure 10 **年龄关于乘客类型箱线图**

由图 9 和图 10 可以看到，比起经济舱超级经济舱，商务舱的年龄更为集中，且偏大。而在乘客类型维度，忠诚客户的年龄比非忠诚的更大。

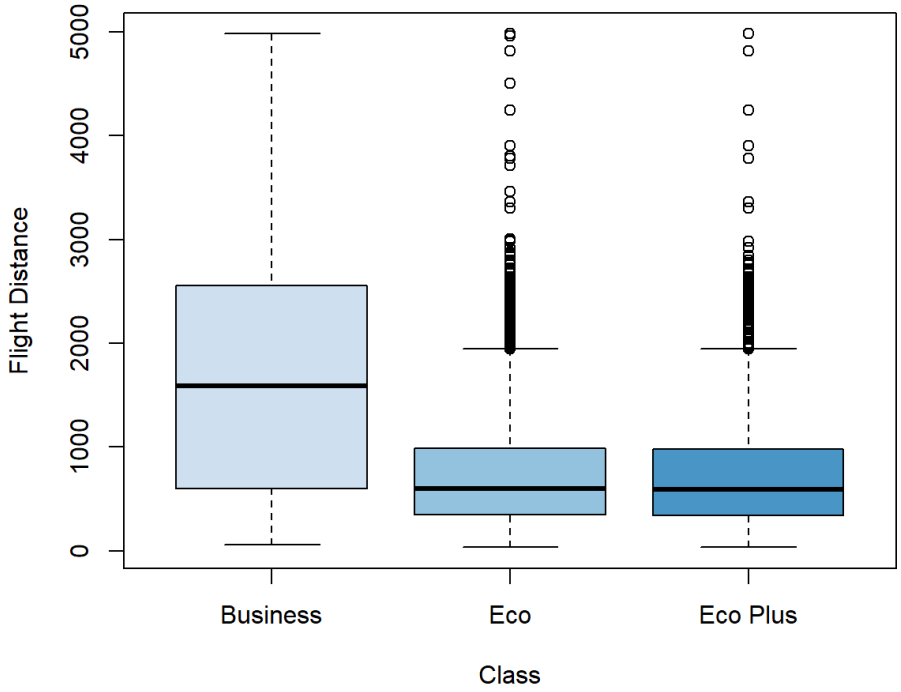

Figure 11 **航程关于舱位箱线图**

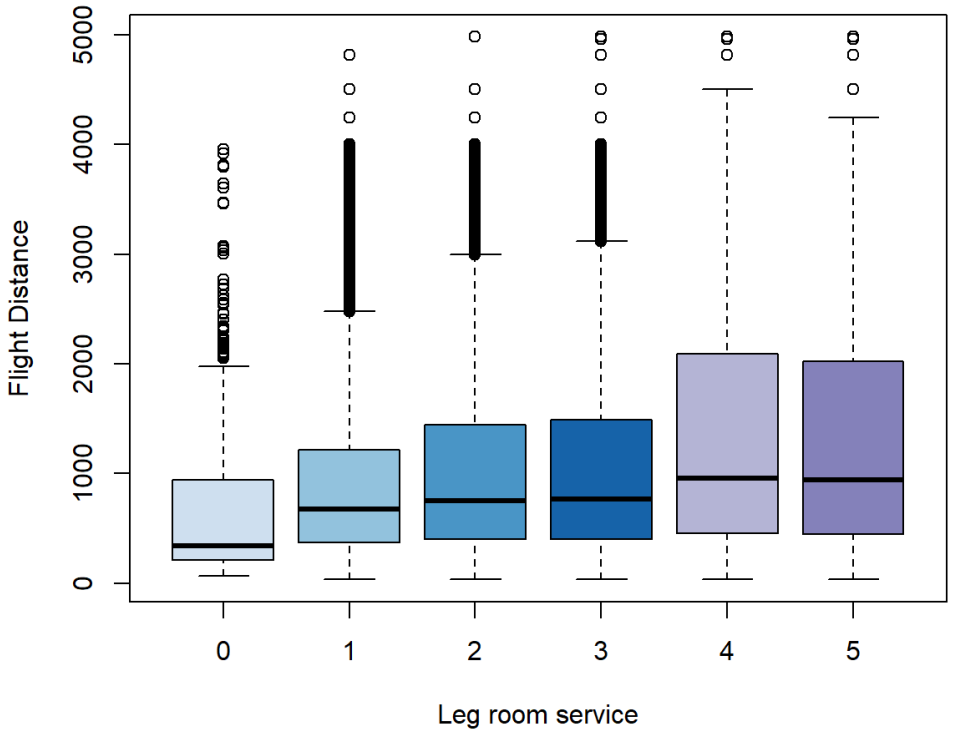

Figure 12 **航行关于腿部服务评分箱线图**

由图 11 可以看到乘坐商务舱的乘客的航行距离比乘坐经济舱的长。由图 12 可以看到长飞行距离往往对应更高的腿部服务评分。

第三章 建模前的准备

在使用分类预测模型评估解释变量对因变量的影响之前，先对数据集进行处理和初步的变量选择。

3.1 数据集缺失值处理

首先对 Arrival Delay in Minutes 这一变量的缺失值进行填充。在 2.2 中，我们已经知道到达延误与出发延误有较强的正相关，因此这里先以到达延误作为因变量，出发延误作为自变量，做一个一元线性回归，再通过得到的线性表达式计算缺失的到达延误时间。

```
Call:
lm(formula = data$Arrival.Delay.in.Minutes ~ data$Departure.Delay.in.Minutes)

Residuals:
    Min      1Q  Median      3Q     Max
-53.510  -1.975  -0.757  -0.461 236.436

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     0.757464   0.029927   25.31   <2e-16 ***
data$Departure.Delay.in.Minutes 0.978849   0.000736 1329.95   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.05 on 129485 degrees of freedom
  (因为不存在，393个观察量被删除了)
Multiple R-squared:  0.9318,    Adjusted R-squared:  0.9318
F-statistic: 1.769e+06 on 1 and 129485 DF,  p-value: < 2.2e-16
```

Figure 13 **到达延误时间~出发延误时间线性回归结果**

如上图，一元线性回归的系数显著性非常强，R 方高达 0.93，因此得到线性关系式，并据此填充到达延误时间的空值：

$$Arrival\ Delay\ in\ Minutes = 0.757464 + 0.978849 * Departure\ Delay\ in\ Minutes$$

3.2 连续型解释变量关于因变量的 t 检验

```
> t.test(data$Age[data$satisfaction == 0],data$Age[data$satisfaction == 1])

        Welch Two Sample t-test

data:  data$Age[data$satisfaction == 0] and data$Age[data$satisfaction == 1]
t = -50.369, df = 129861, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4.249171 -3.930866
sample estimates:
mean of x mean of y
 37.65100  41.74102

> t.test(data$Flight.Distance[data$satisfaction == 0],data$Flight.Distance[data$satisfaction == 1])

        Welch Two Sample t-test

data:  data$Flight.Distance[data$satisfaction == 0] and data$Flight.Distance[data$satisfaction == 1]
t = -107.63, df = 96579, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -610.7466 -588.9009
sample estimates:
mean of x mean of y
 929.7154 1529.5392

> t.test(data$Departure.Delay.in.Minutes[data$satisfaction == 0],data$Departure.Delay.in.Minutes[data$satisfaction == 1])

        Welch Two Sample t-test

data:  data$Departure.Delay.in.Minutes[data$satisfaction == 0] and data$Departure.Delay.in.Minutes[data$satisfaction == 1]
t = 18.641, df = 127842, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 3.487307 4.306802
sample estimates:
mean of x mean of y
 16.40684  12.50978
> t.test(data$Arrival.Delay.in.Minutes[data$satisfaction == 0],data$Arrival.Delay.in.Minutes[data$satisfaction == 1])

        Welch Two Sample t-test

data:  data$Arrival.Delay.in.Minutes[data$satisfaction == 0] and data$Arrival.Delay.in.Minutes[data$satisfaction == 1]
t = 21.287, df = 127685, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 4.096622 4.927522
sample estimates:
mean of x mean of y
 17.12030  12.60822
```

Figure 14 **连续型解释变量 t 检验结果**

为了探究连续型解释变量与因变量的关系，并初步对放入模型的变量做选择，分别做 t 检验，结果如图 14 所示。4 项检验均显著，即满意/中立或不满意的乘客在年龄、航程、出发延误时间、到达延误时间上有显著差异。因此，建模时这

四个变量都需要放到模型中（由于具有线性关系，出发延误时间、到达延误时间可选其一）。

3.3 属性解释变量与因变量的列联表分析

接下来对各属性变量做与因变量的列联表分析，检验其独立性。以下以性别为例，解释结果。原假设是两个变量之间存在独立性，在 H0 下得到各格的期望值，如表格 2 所示。

表格 2 satisfaction*gender **二维列联表**

	Female	Male	Sum
Neutral or dissatisfied	37630 (37268.35)	35822 (36183.65)	73452
Satisfied	28269 (28630.65)	28159 (27797.35)	56428
Sum	65899	63981	129880

可以用卡方统计量进行检验，这里选取X^2，统计量近似卡方分布，计算得到：

$$X^2 = \sum_i \sum_j (n_{ij} - \mu_{ij})^2 / \mu_{ij} = 16.3974, p - value = 5.13559e - 05$$

拒绝原假设，即不能认为性别与满意与否独立。

此外，还可以对相对风险比和发生比进行统计检验。得到 Risk Ratio 估计值为 1.03，95%置信区间为(1.02,1.04)；Odds Ratio 估计值为 1.05，95%置信区间为(1.02,1.07)。因此，我们可以推断出女性对航程满意的概率为男性的 1.02~1.04 倍。

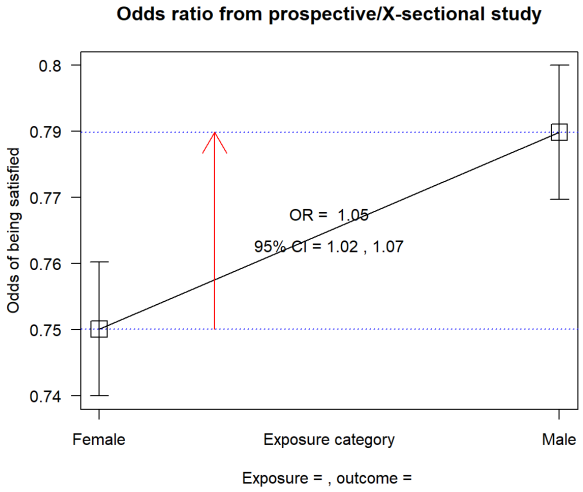

Figure 15 satisfaction*gender **列联表** Odds Ratio **检验**

由于篇幅限制，这里不一一详细解释列联表的结果，18 个属性解释变量与因变量的列联表结果归纳如下表。

表格 3 **二维列联表结果一览**

Satisfaction *	X^2	$p-value$	认为独立
Gender	16.3974	<0.001	否
Customer Type	4494.159	<0.001	否
Type of Travel	26284.48	<0.001	否
Class	32906.17	<0.001	否
Inflight wifi service	35891.43	<0.001	否
Departure/Arrival time convenient	601.463	<0.001	否
Ease of Online booking	12846.7	<0.001	否
Gate location	3069.908	<0.001	否
Food and drink	6571.203	<0.001	否
Online boarding	49531.22	<0.001	否
Seat comfort	19538.74	<0.001	否
Inflight entertainment	23071.6	<0.001	否
On-board service	14342.66	<0.001	否
Leg room service	15200.78	<0.001	否
Baggage handling	10820.21	<0.001	否
Checkin service	8143.773	<0.001	否

Inflight service	10357.93	<0.001	否
Cleanliness	12948.92	<0.001	否

由表中结果，各属性解释变量与因变量的相关性均不能排除，后续建模时都应该纳入模型中。

第四章 旅途满意度的影响因素研究

4.1 基于逻辑回归的满意度影响因素研究

一、模型初步构建

首先，将各属性变量全部转变为 0-1 变量，在训练集上训练逻辑回归模型，结果如下图所示。

```
Coefficients: (3 not defined because of singularities)
                                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                             4.447e+00  9.961e+03   0.000 0.999644
Gender                                 -4.657e-02  2.730e-02  -1.706 0.088027 .
Customer.Type                          -3.355e+00  4.953e-02 -67.728  < 2e-16 ***
Age                                    -2.307e-03  1.017e-03  -2.269 0.023264 *
Type.of.Travel                          4.272e+00  5.507e-02  77.582  < 2e-16 ***
Flight.Distance                         7.075e-06  1.535e-05   0.461 0.644914
Departure.Delay.in.Minutes              5.117e-03  1.328e-03   3.854 0.000116 ***
Arrival.Delay.in.Minutes               -8.939e-03  1.309e-03  -6.829 8.54e-12 ***
Eco                                     2.067e-01  5.840e-02   3.540 0.000400 ***
Business                                8.363e-01  6.049e-02  13.826  < 2e-16 ***
Inflight.wifi.service.1                -2.402e+01  8.867e+01  -0.271 0.786512
Inflight.wifi.service.2                -2.427e+01  8.867e+01  -0.274 0.784278
Inflight.wifi.service.3                -2.432e+01  8.867e+01  -0.274 0.783885
Inflight.wifi.service.4                -2.277e+01  8.867e+01  -0.257 0.797372
Inflight.wifi.service.5                -1.720e+01  8.867e+01  -0.194 0.846228
Departure.Arrival.time.convenient.1     3.146e-01  9.296e-02   3.384 0.000714 ***
Departure.Arrival.time.convenient.2     4.307e-01  8.959e-02   4.807 1.53e-06 ***
Departure.Arrival.time.convenient.3     2.420e-01  8.632e-02   2.804 0.005051 **
Departure.Arrival.time.convenient.4    -6.775e-01  7.733e-02  -8.761  < 2e-16 ***
Departure.Arrival.time.convenient.5    -9.132e-01  8.492e-02 -10.754  < 2e-16 ***
Ease.of.Online.booking.1                3.066e+00  9.145e-01   3.353 0.000800 ***
Ease.of.Online.booking.2                2.998e+00  9.145e-01   3.278 0.001045 **
Ease.of.Online.booking.3                3.499e+00  9.143e-01   3.827 0.000130 ***
Ease.of.Online.booking.4                4.344e+00  9.141e-01   4.753 2.01e-06 ***
Ease.of.Online.booking.5                3.713e+00  9.144e-01   4.061 4.89e-05 ***
Gate.location.1                        -1.876e+01  6.523e+03  -0.003 0.997705
Gate.location.2                        -1.868e+01  6.523e+03  -0.003 0.997715
Gate.location.3                        -1.885e+01  6.523e+03  -0.003 0.997695
Gate.location.4                        -1.910e+01  6.523e+03  -0.003 0.997663
Gate.location.5                        -1.931e+01  6.523e+03  -0.003 0.997638
Food.and.drink.1                       -3.197e-01  1.750e+00  -0.183 0.855020
Food.and.drink.2                       -3.719e-02  1.749e+00  -0.021 0.983041
Food.and.drink.3                       -1.676e-01  1.749e+00  -0.096 0.923680
Food.and.drink.4                       -1.223e-01  1.749e+00  -0.070 0.944275
Food.and.drink.5                       -2.766e-01  1.749e+00  -0.158 0.874355
Online.boarding.1                      -3.625e+00  9.181e-01  -3.948 7.88e-05 ***
Online.boarding.2                      -3.545e+00  9.180e-01  -3.861 0.000113 ***
Online.boarding.3                      -3.776e+00  9.177e-01  -4.115 3.87e-05 ***
Online.boarding.4                      -2.130e+00  9.174e-01  -2.322 0.020218 *
Online.boarding.5                      -8.813e-01  9.176e-01  -0.960 0.336838
Seat.comfort.1                          2.047e+01  6.523e+03   0.003 0.997497
Seat.comfort.2                          1.994e+01  6.523e+03   0.003 0.997561
Seat.comfort.3                          1.888e+01  6.523e+03   0.003 0.997690
Seat.comfort.4                          1.959e+01  6.523e+03   0.003 0.997604
Seat.comfort.5                          2.043e+01  6.523e+03   0.003 0.997501
Inflight.entertainment.1                3.969e+01  1.516e+03   0.026 0.979106
Inflight.entertainment.2                4.045e+01  1.516e+03   0.027 0.978709
Inflight.entertainment.3                4.128e+01  1.516e+03   0.027 0.978270
Inflight.entertainment.4                4.095e+01  1.516e+03   0.027 0.978443
Inflight.entertainment.5                4.019e+01  1.516e+03   0.027 0.978845
On.board.service.1                     -2.334e+01  4.052e+03  -0.006 0.995403
On.board.service.2                     -2.319e+01  4.052e+03  -0.006 0.995433
On.board.service.3                     -2.266e+01  4.052e+03  -0.006 0.995537
On.board.service.4                     -2.258e+01  4.052e+03  -0.006 0.995554
On.board.service.5                     -2.204e+01  4.052e+03  -0.005 0.995659
Leg.room.service.1                     -2.401e+00  9.585e-01  -2.505 0.012262 *
Leg.room.service.2                     -2.127e+00  9.580e-01  -2.221 0.026383 *
Leg.room.service.3                     -2.244e+00  9.579e-01  -2.342 0.019158 *
Leg.room.service.4                     -1.545e+00  9.580e-01  -1.613 0.106738
Leg.room.service.5                     -1.384e+00  9.577e-01  -1.445 0.148583
Baggage.handling.2                     -2.198e-01  7.600e-02  -2.892 0.003826 **
Baggage.handling.3                     -8.434e-01  7.098e-02 -11.882  < 2e-16 ***
Baggage.handling.4                     -2.451e-01  6.901e-02  -3.552 0.000382 ***
Baggage.handling.5                      5.159e-01  7.336e-02   7.033 2.03e-12 ***
Checkin.service.1                      -1.426e+00  5.429e-02 -26.265  < 2e-16 ***
Checkin.service.2                      -1.235e+00  5.402e-02 -22.863  < 2e-16 ***
Checkin.service.3                      -7.256e-01  4.346e-02 -16.696  < 2e-16 ***
Checkin.service.4                      -7.451e-01  4.324e-02 -17.231  < 2e-16 ***
Checkin.service.5                              NA         NA      NA       NA
Inflight.service.1                     -4.820e-01  7.644e-02  -6.305 2.88e-10 ***
Inflight.service.2                     -7.008e-01  6.932e-02 -10.110  < 2e-16 ***
Inflight.service.3                     -1.395e+00  5.729e-02 -24.345  < 2e-16 ***
Inflight.service.4                     -6.952e-01  4.493e-02 -15.472  < 2e-16 ***
Inflight.service.5                             NA         NA      NA       NA
Cleanliness.1                          -9.978e-01  7.512e-02 -13.283  < 2e-16 ***
Cleanliness.2                          -9.549e-01  7.304e-02 -13.074  < 2e-16 ***
Cleanliness.3                          -4.678e-01  6.144e-02  -7.613 2.68e-14 ***
Cleanliness.4                          -6.023e-01  6.022e-02 -10.003  < 2e-16 ***
Cleanliness.5                                  NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 16 **初始逻辑回归模型结果**

根据图 16 的回归模型拟合结果，显著性强的变量和几乎无关的变量数量对半开，且全部变量共有 90 个左右，一起纳入模型会使得模型实在太臃肿。因此，下一步进行变量筛选。

二、模型改善

为了给模型降维，首先选取初始模型中显著性较强的自变量，再通过 stepwise 方法进行变量选择。得到所需的自变量为：性别、乘客类型、年龄、出行目的、出发延误时间、是否经济舱、是否商务舱、出发时间便捷性满意度（水平 1-5）、网上订票简易性满意度（水平 1-5）、线上值机满意度（水平 1-5）、腿

部服务满意度（水平 1-5）、行李处理满意度（水平 2、3、5）、登机服务满意度（水平 1-4）、航行中服务满意度（水平 1-5）、整洁度（水平 1-5）。

三、拟合结果与分析

```
Coefficients:
                                   Estimate Std. Error z value Pr(>|z|)
(Intercept)                       -2.665e+01  1.928e+02  -0.138   0.8900
Gender                            -3.986e-02  2.243e-02  -1.777   0.0755 .
Customer.Type                     -2.598e+00  3.623e-02 -71.700  < 2e-16 ***
Age                               -5.009e-03  8.120e-04  -6.168 6.91e-10 ***
Type.of.Travel                     3.610e+00  3.952e-02  91.334  < 2e-16 ***
Departure.Delay.in.Minutes        -4.476e-03  2.928e-04 -15.287  < 2e-16 ***
Eco                                1.833e-01  4.508e-02   4.065 4.79e-05 ***
Business                           7.017e-01  4.492e-02  15.620  < 2e-16 ***
Departure.Arrival.time.convenient.1  4.130e-01  6.843e-02   6.035 1.59e-09 ***
Departure.Arrival.time.convenient.2  5.838e-01  6.604e-02   8.840  < 2e-16 ***
Departure.Arrival.time.convenient.3  4.127e-01  6.441e-02   6.408 1.48e-10 ***
Departure.Arrival.time.convenient.4 -4.421e-01  5.779e-02  -7.651 1.99e-14 ***
Departure.Arrival.time.convenient.5 -6.718e-01  6.055e-02 -11.096  < 2e-16 ***
Ease.of.Online.booking.1          -3.474e+00  1.226e-01 -28.332  < 2e-16 ***
Ease.of.Online.booking.2          -3.771e+00  1.213e-01 -31.083  < 2e-16 ***
Ease.of.Online.booking.3          -3.467e+00  1.200e-01 -28.896  < 2e-16 ***
Ease.of.Online.booking.4          -2.259e+00  1.173e-01 -19.250  < 2e-16 ***
Ease.of.Online.booking.5          -1.770e+00  1.177e-01 -15.035  < 2e-16 ***
Online.boarding.1                 -6.952e-01  1.264e-01  -5.500 3.79e-08 ***
Online.boarding.2                 -8.446e-01  1.261e-01  -6.696 2.14e-11 ***
Online.boarding.3                 -1.022e+00  1.251e-01  -8.168 3.14e-16 ***
Online.boarding.4                  1.063e+00  1.244e-01   8.547  < 2e-16 ***
Online.boarding.5                  2.971e+00  1.268e-01  23.424  < 2e-16 ***
Leg.room.service.1                 7.165e-01  1.734e-01   4.132 3.60e-05 ***
Leg.room.service.2                 1.063e+00  1.720e-01   6.184 6.25e-10 ***
Leg.room.service.3                 1.137e+00  1.717e-01   6.618 3.65e-11 ***
Leg.room.service.4                 2.044e+00  1.720e-01  11.885  < 2e-16 ***
Leg.room.service.5                 2.124e+00  1.727e-01  12.303  < 2e-16 ***
Baggage.handling.2                -3.717e-01  4.431e-02  -8.389  < 2e-16 ***
Baggage.handling.3                -6.148e-01  3.320e-02 -18.519  < 2e-16 ***
Baggage.handling.5                 7.147e-01  3.280e-02  21.790  < 2e-16 ***
Checkin.service.1                 -1.241e+00  4.324e-02 -28.688  < 2e-16 ***
Checkin.service.2                 -1.092e+00  4.288e-02 -25.457  < 2e-16 ***
Checkin.service.3                 -6.581e-01  3.526e-02 -18.664  < 2e-16 ***
Checkin.service.4                 -6.983e-01  3.506e-02 -19.915  < 2e-16 ***
Inflight.service.1                 1.244e+01  1.730e+02   0.072   0.9427
Inflight.service.2                 1.249e+01  1.730e+02   0.072   0.9425
Inflight.service.3                 1.227e+01  1.730e+02   0.071   0.9435
Inflight.service.4                 1.303e+01  1.730e+02   0.075   0.9400
Inflight.service.5                 1.361e+01  1.730e+02   0.079   0.9373
Cleanliness.1                      1.196e+01  8.496e+01   0.141   0.8881
Cleanliness.2                      1.217e+01  8.496e+01   0.143   0.8861
Cleanliness.3                      1.246e+01  8.496e+01   0.147   0.8834
Cleanliness.4                      1.269e+01  8.496e+01   0.149   0.8812
Cleanliness.5                      1.309e+01  8.496e+01   0.154   0.8776
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 17 **逻辑回归结果**

根据图 17 的回归结果可知，乘客类型、年龄、出行原因、出发延误时间、舱位、出发到达时间是否便利、网上订票是否容易、线上值机、腿部服务、行李处理、登机服务对乘客满意度的影响较为显著。其中，忠实乘客、年轻乘客、商务出行都会提高乘客对行程满意的概率，这是符合经验和直觉的。航空公司应该在这些方面加以研究和提高，提升乘客的满意程度。

四、模型评价与预测表现

将测试集上训练出的模型用于测试集，根据自变量对满意结果做预测，再与测试集本身满意情况比较，得到结果如下。

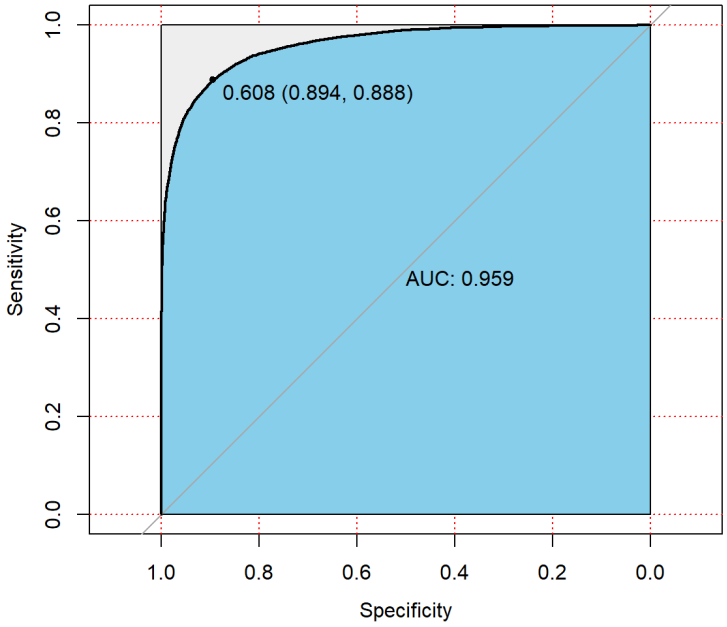

Figure 18 逻辑回归模型在测试集上的 ROC 曲线

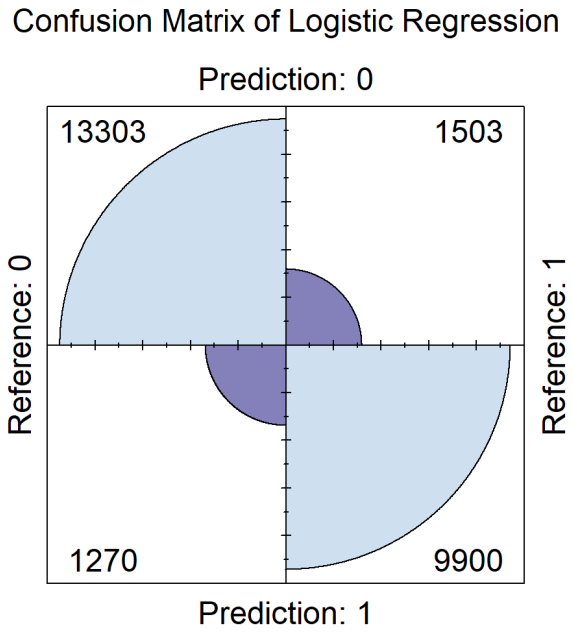

Figure 19 逻辑回归模型预测混淆矩阵

由图 18 可知，改良后的逻辑回归模型在测试集上达到了 0.959 的 AUC 值，拟合效果非常好。由图 19 可知，模型达到了 89.32%的准确率、86.82%的召回率、

91.29%特异性，预测效果也非常好。

4.2 基于其他流行分类算法的研究

除了逻辑回归，这一部分希望通过其他流行的分类算法来建模分析，这里主要采用 KNN 最近邻算法和随机森林两种方法。

一、KNN 最近邻算法

（一）模型介绍

KNN 最近邻算法是最简单的机器学习方法，它的思路非常简单直观：如果一个样本在特征空间中的 K 个最相似（即特征空间中最邻近）的样本中的大多数属于某一个类别，则该样本也属于这个类别。该方法在定类决策上只依据最邻近的一个或者几个样本的类别来决定待分样本所属的类别。

总体来说，KNN 分类算法包括以下 4 个步骤：

- 通过交叉验证计算方差，确定合适的 K 值；
- 根据公式$d(x,y) := \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$计算待分类点到其他每个样本点的距离；
- 对每个距离进行排序，然后选择出距离最小的 K 个点；
- 对 K 个点所属的类别进行比较，将测试样本点归入在 K 个点中占比最高的那一类。

（二）模型评价与预测表现

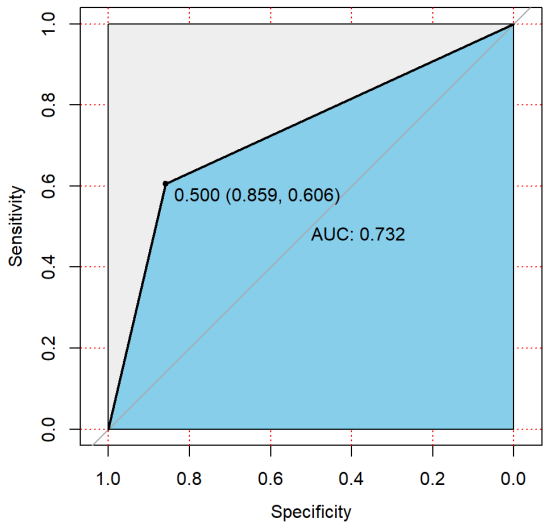

Figure 20 KNN 最近邻算法在测试集上的 ROC 曲线

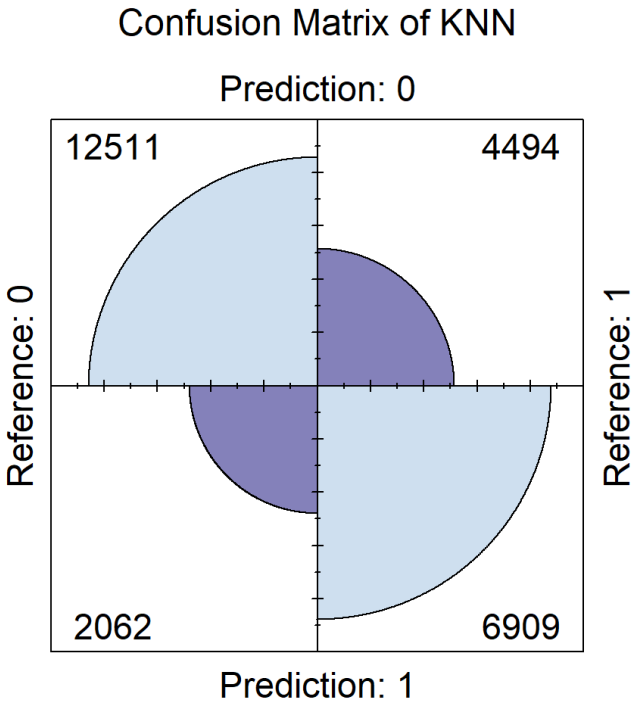

Figure 21 KNN **最近邻算法预测的混淆矩阵**

由图 20 可知，KNN 最近邻算法的拟合程度不佳，AUC 值仅有 0.732，图 21 中的混淆矩阵也体现分别仅有 71.61%、60.6%的准确率和召回率。因此 KNN 在本数据集上的预测表现不如逻辑回归好。

二、随机森林

（一）模型介绍

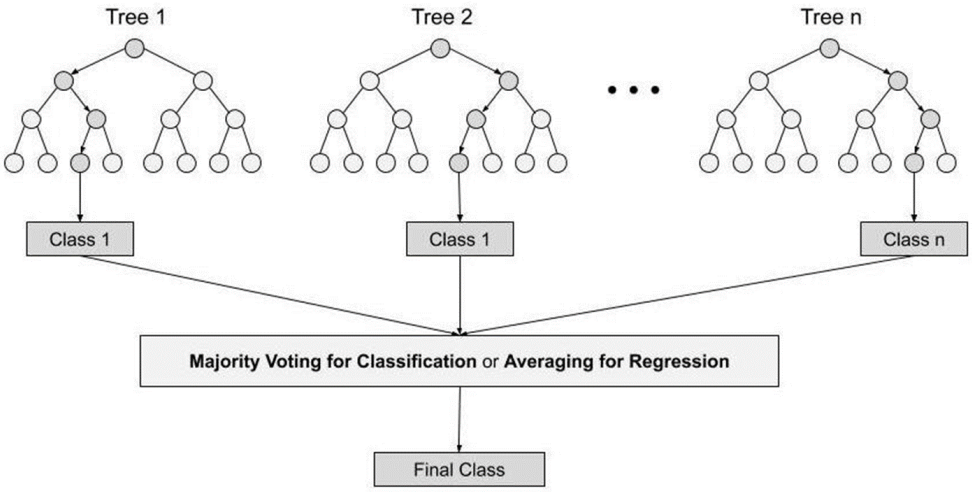

Figure 22 **随机森林算法图示**

随机森林是通过随机的方式形成多棵决策树，并通过所有决策树的投票结果选出最终预测结果的一种 Bagging 集成算法，其本质是利用多个弱分类器结合在

一起，形成更稳定的强分类器。

（二）模型评价与预测表现

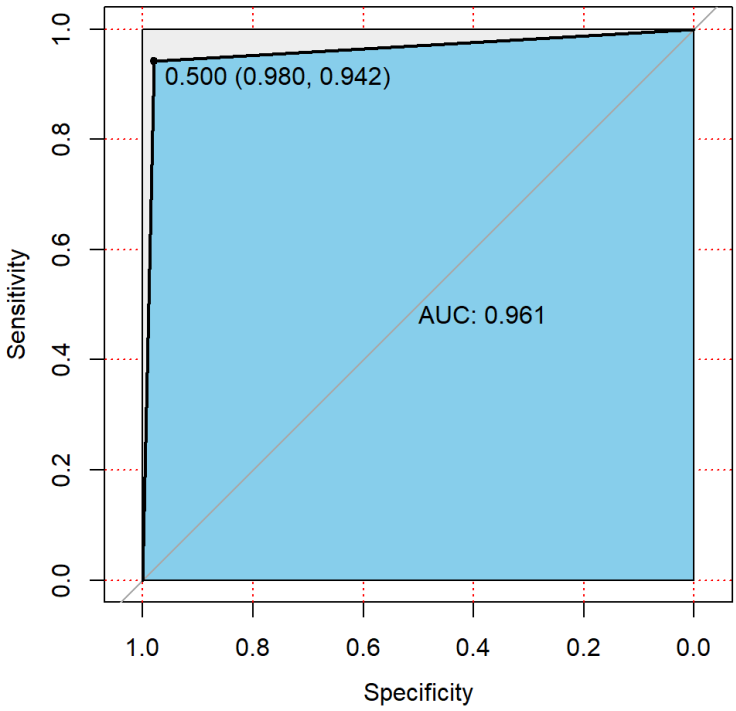

Figure 23 随机森林算法在测试集上的 ROC 曲线

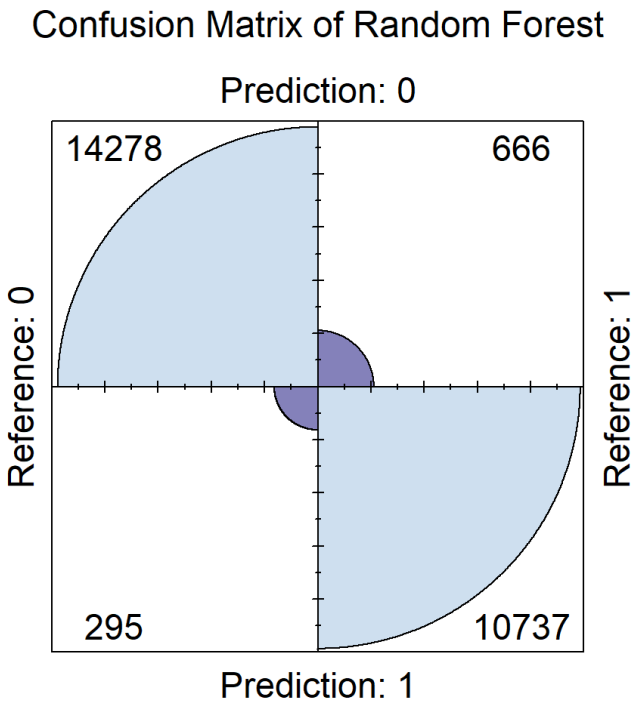

Figure 24 随机森林算法预测的混淆矩阵

由图 23、图 24 可以看到，随机森林在测试集上的 AUC 值高达 0.961，预测准确率、召回率、特异性分别为 96.3%、94.16%、97.98%。模型拟合效果略优于逻辑回归，预测效果明显优于逻辑回归。

另外，根据随机森林的变量重要性分析，在模型中解释性最强的两个变量为出行类型、舱位，这与描述性统计和逻辑回归中得到的结果吻合。也就是说，商务出行、商务舱对提高满意概率的作用比较显著。

第五章 结论与建议

5.1 结论及不足

根据前文的分析，得到以下结论：

1. 长距离的航班满意度更高
2. 乘客类型、年龄、出行原因、出发延误时间、舱位、出发到达时间是否便利、网上订票是否容易、线上值机、腿部服务、行李处理、登机服务对乘客满意度的影响较为显著；
3. 忠实乘客、年轻乘客、商务出行都会提高乘客对行程满意的概率；
4. 是否商务出行和舱位是重要性最大的两个变量。

此外，在分析中，有以下值得改进的地方：

1. 建模求解时，虽然筛选了一些变量，但最终纳入模型的变量还是比较多，尤其是乘客评分类型的自变量，数量多且较杂乱。可以尝试利用主成分分析、因子分析等方法对变量进行进一步的分类和降维，且归纳出每一类自变量的意义；
2. 在第四章中独立训练了三个机器学习模型，其中随机森林的拟合和预测效果很不错，但另外两个依然有优化的余地。可以尝试模型融合等方式，在这三个模型的基础之上训练出效果更好的模型。

5.2 建议

结合本文的分析，为航空公司提出以下运营建议：

1. 在乘客层面，如果航空公司的策略是巩固满意度已经较高的乘客，应该着重发展 40 岁以下的年轻乘客，以及该公司的忠诚客户，可以针对他

们特别制定一些促销活动，或是机上装潢，餐饮选择等；反之，如果策略是发展满意度还不够高的乘客，应该主要针对 40 岁以上的乘客，或非重复乘坐航班的乘客，可以改善机上设施，变得更加老人友好，或展开一些新人入会的优惠活动；

2. 在航班安排层面，可以在商务出行比较集中的航班增加商务舱席位，并在业界的竞争中着重抢夺出发/到达时间便利、舒适的航线；
3. 在服务层面，航空公司可以着重提高线上平台的服务，比如网上订票、线上值机、航班晚点提醒等，让旅客的出行更为便利，从而提升满意程度。此外，距离较长的航线，长时间坐在狭小空间可能导致腿部麻木、水肿，而腿部空间是乘客在评价一段旅途时比较看重的因素，因此，在新型号客机的设计上，可以更加注重座椅腿部空间的设计。