基础知识

基本方法

实验分析

实验数据描述

我们在kaggle官网上下载了官方的数据集进行实验。这个数据集大约有13000个样本，13个特征参数，1个标签（待预测项）。这13个参数均会对乘客的去向产生影响。数据集分为训练集和测试集合，训练集包含大约三分之二（~8700）的数据（train.csv），测试集包含大约三分之一（~4300）的数据（test.csv）。这些数据中，每一个数据元素包含14个属性，1个是标签（即是否被传送到其他星球），13个是特征参数，均会对结果产生影响。其中，

passangerID是每一位乘客的ID，每个ID的形式都表示乘客与一个团体一起旅行，小组的成员通常是家庭成员。比如0013\_01，表示是0013组01号成员。

HomePlanet是指乘客离开的星球，通常是他们永久居住的星球。

CryoSleep指的是乘客是否在航行期间保持假死的状态。处于假死状态的乘客将被限制在他们的客舱内。

Cabin指的是乘客所住的仓号。采用的形式是deck/num/side(P/S)。

Destination表示乘客将要前往的星球。

Age表示乘客的年龄。

VIP表示乘客是否为VIP(TRUE/FALSE)。

RoomService, FoodCourt, ShoppingMall. Spa表示乘客在飞船上许多豪华设施中为购买服务交纳的金额。

Name表示乘客的名字。

Transported表示乘客是否被传送到另一个维度。

缺失值和数据类型分析

为了保证训练的准确性，首先要做的便是对训练数据集的缺失值分析和数据类型分析。

使用isna().sum()函数统计出，整个训练数据集中一共有2324个缺失的数据。针对每一个特征参数进行分析，得出的结果如下表所示：

|  |  |  |  |
| --- | --- | --- | --- |
| 属性名称 | 缺失值数量 | 属性名称 | 缺失值数量 |
| CryoSleep | 217 | ShoppingMall | 208 |
| VIP | 203 | HomePlanet | 201 |
| Name | 200 | Cabin | 199 |
| VRDeck | 188 | FoodCourt | 183 |
| Spa | 183 | Destination | 182 |
| RoomService | 181 | Age | 179 |
| PassengerId | 0 | Transported | 0 |

表1 各属性缺失值数量

基本所有的属性均含有缺失值，所以处理缺失值是一个至关重要的问题。

接着本文又对数据集进行了重复值检查，幸运的是，所有的训练集和测试集均没有重复的数据。

针对实验数据的描述我们可以发现，类似于PassangerID这种属于字符型变量，应该是属于描述性属性，而对于金额这一类数据，应该是连续性浮点型或者整形，对于他们的处理方式也该是不同的，所以本文还做了如下的分析：

首先对于每一种属性进行唯一值统计：结果如下表所示：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 唯一值数量范围 | [1,10] | (10,100] | (100,8000] | 大于8000 |
| 属性名称 | HomePlanet, CryoSleep, Destination, VIP，Transported | Age | Cabin, RoomService, FoodCourt,  ShoppingMall, Spa, VRDeck | PassengerId, Name |

表2 各属性的唯一值数量

对于每一个属性的属性类型，如下表所示：

|  |  |
| --- | --- |
| object | PassengerId, HomePlanet, CryoSleep, Cabin, Destination, VIP, Name |
| float64 | Age, RoomService, FoodCourt, ShoppingMall, Spa, VRDeck, |
| bool | Transported |

表3 各属性的属性类型

对于属性不同的值仅有1-10的属性，显然他们类别特征属性的变量，即他们的取值种类数目特别少，像是否选择假死睡眠、VIP、是否被传送，只有是和否两种选择；而对于超过8000的属性，即每条记录有一个不同的属性，即每个人独一无二的属性，类似于人的名字和乘客编号。而对于年龄，用户的付费服务（RoomService，FoodCourt，ShoppingMall，Spa，VRdeck）这些数据记录的是连续的数值型，年龄较为特殊，都是整数且均在0到80之间。这里最需要注意的是甲板（Deck）类型，这里分成三组数据，分别是deck/num/side，deck表示甲板号，num表示编号，side表示两侧（仅有两个取值），对于这个数据处理要采用特别的方式。数据的更详细的分析如下文所示。由于姓名，乘客编号和仓号这三个参数可以唯一确定一个乘客，所以我们称之为定性特征。

探索性数据分析

由于这次实验的目的是需要预测乘客是否被传送到另外一个星系，即transported的取值，所以对于transported的分布有一定的要求，如果数据分布的不平很还需要采取上采样或者下采样的策略。如下图是对transported的分布规律的探索：

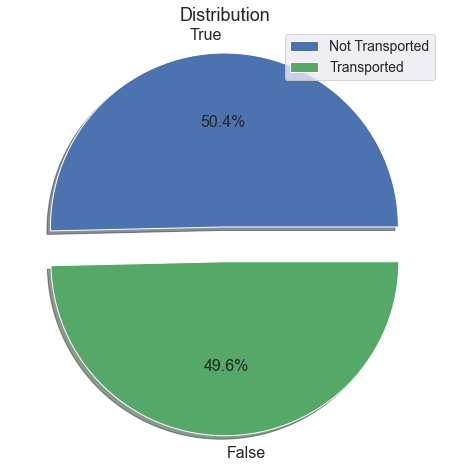


图1 transported的分布

可以发现，transported基本是对半分的，所以这个数据集是比较均衡的，不需要进行上采样或者下采样的处理。

接下来对于各个属性进行单独分析。

连续性变量

首先是年龄。由第一步分数据分析可知，年龄的不同值数量仅有80中不同的取值且均为整数，所以可以使用直方图图来分析连续变量年龄的分布情况。直方图如下所示：

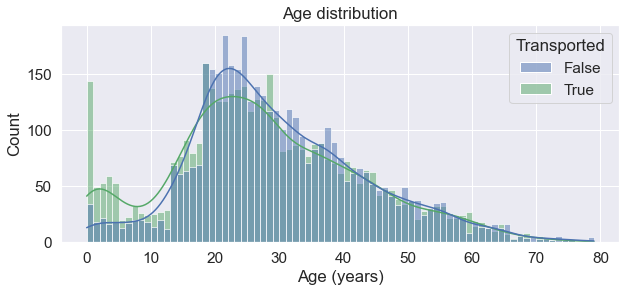


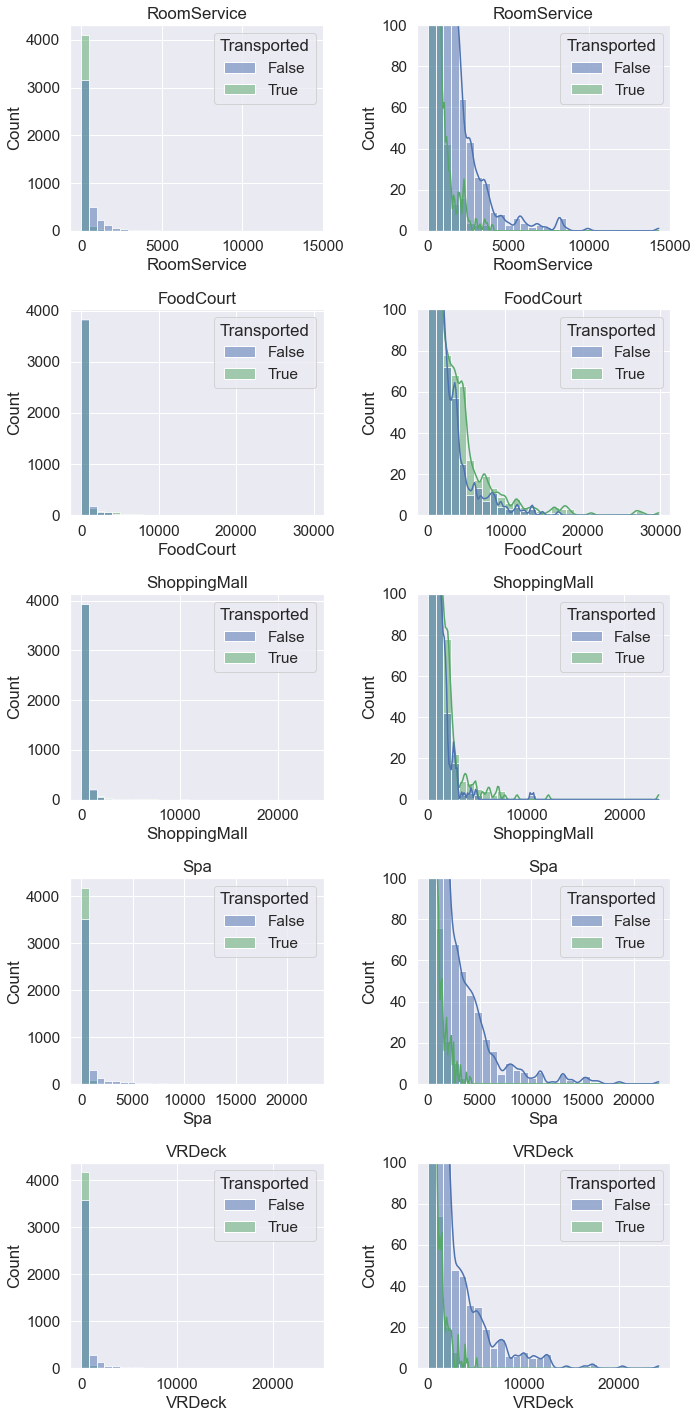
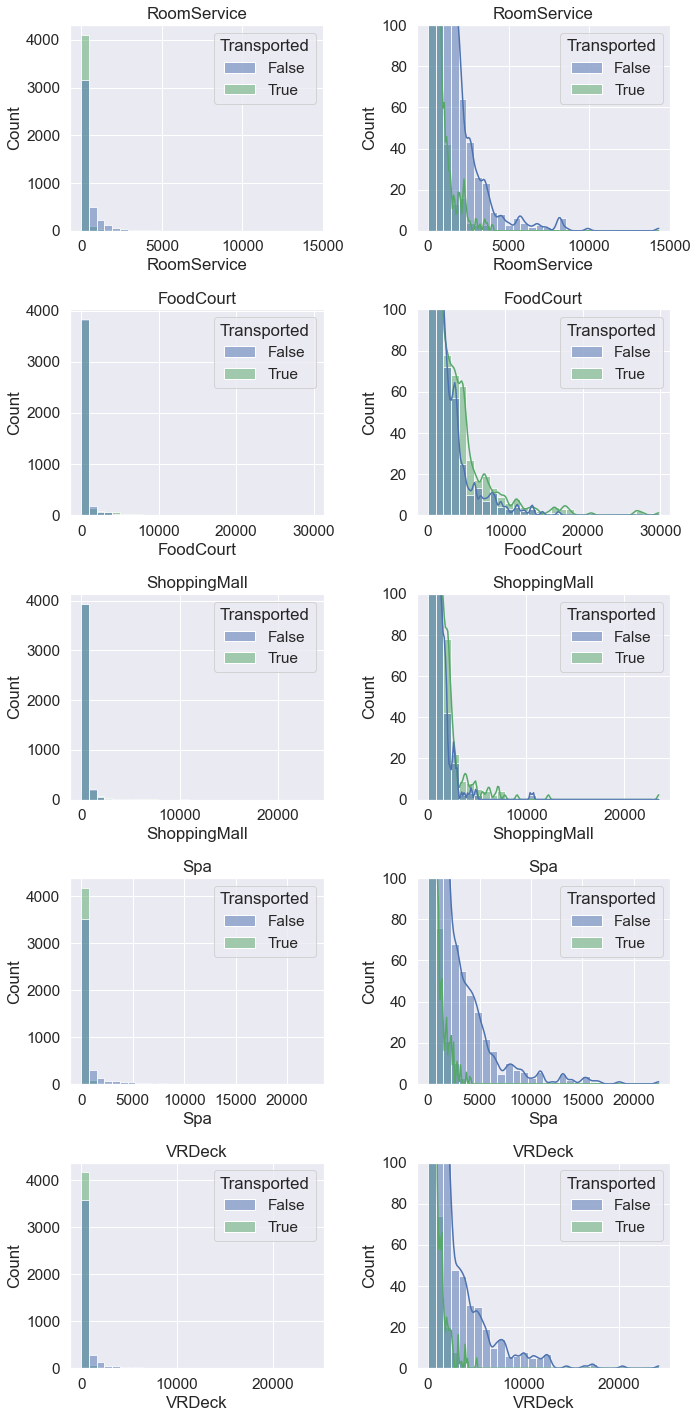
图2 年龄直方图

从中可以得出以下结论：

* 0-18岁的乘客中，传送成功的比例更高
* 18-25岁的乘客中，传送成功的比例更低
* 25岁以上的乘客中，传送成功的和不成功的比例基本相同

这给到我们一个启示可以按照三个年龄段对于数据进行划分，这样可以更好的进行模型的训练。

接下来是对各个付费服务属性（RoomService，FoodCourt，ShoppingMall，Spa，VRdeck）的分析。由于各个花费是连续性的变量，所以仍然可以使用直方图的形式行进分析。以下是直方图：



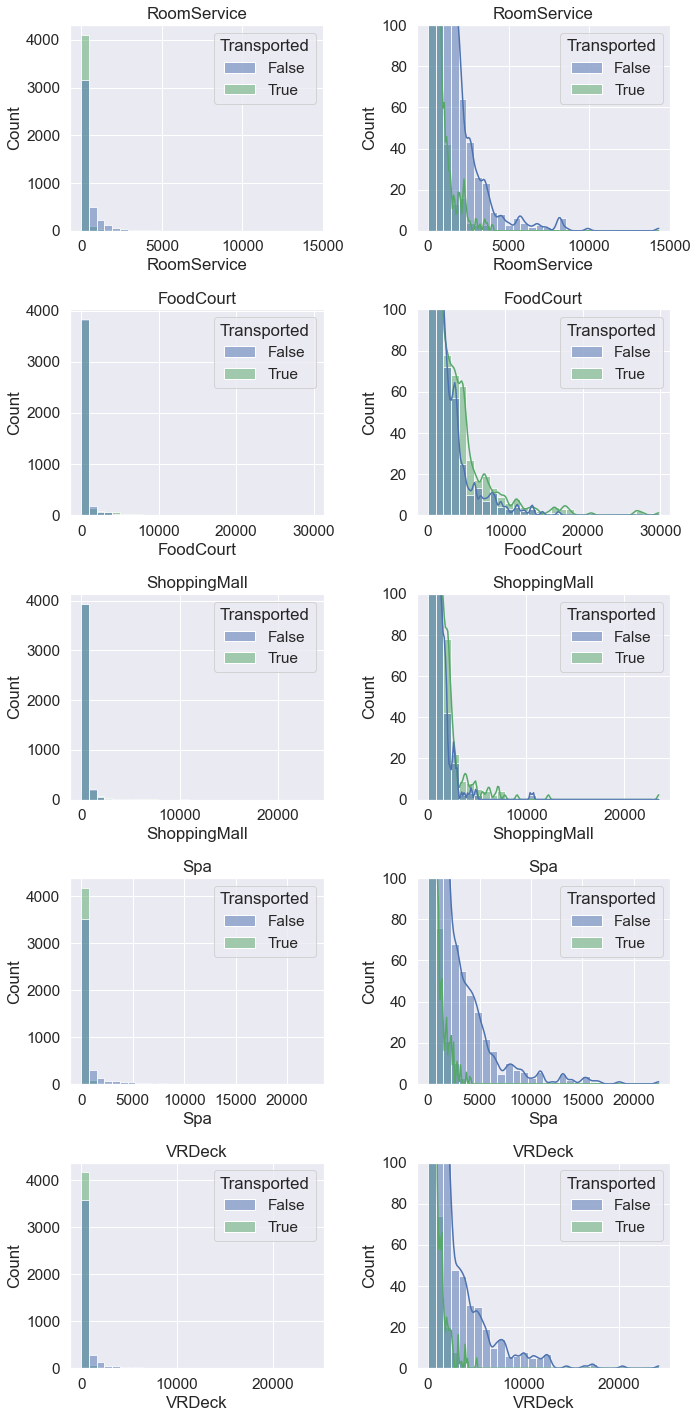


图3 付费服务属性直方图

从中我们可以得到下面的结论：

* 大多数的人没有花多少钱
* 花费的分布呈指数衰减（如图3所示）
* 存在一小部分异常值
* 被传送的人倾向于花费较少

RoomService（客房服务）、Spa（温泉）和VRDeck（虚拟现实甲板）与FoodCourt（美食广场）和ShoppingMall（购物中心）有不同的分布 - 我们可以将其视为奢侈品与基本设施。

这些在数据处理当中可以给我们下面的启示：

* 创建一个新的特征，跟踪所有5个设施的总支出。
* 创建一个二进制特征，用于指示该人是否没有花费任何金额（即总支出为0）。
* 进行对数转换以减小偏度。

类别特征变量

经第一步分析得，'HomePlanet', 'CryoSleep', 'Destination', 'VIP'，这些属性属于类别特征变量。接着用可视化的方法，生成每个分类特征的计数图，其中x轴表示特征中的类别，y轴表示每个类别中的观测计数。这些计数图如下所示：

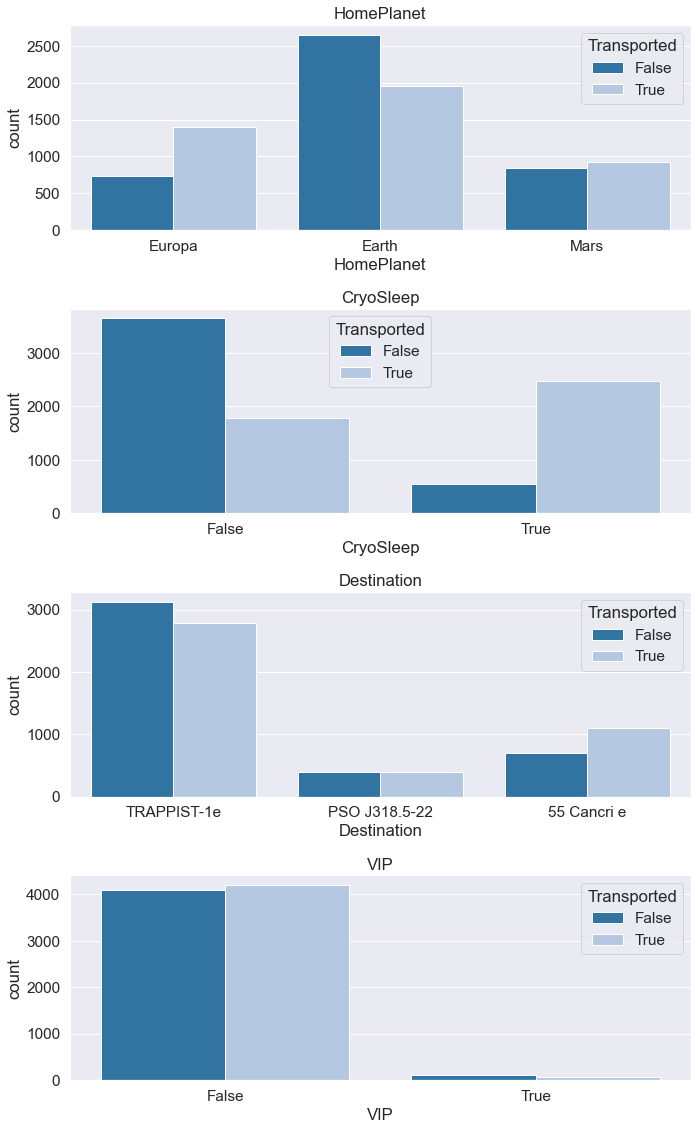
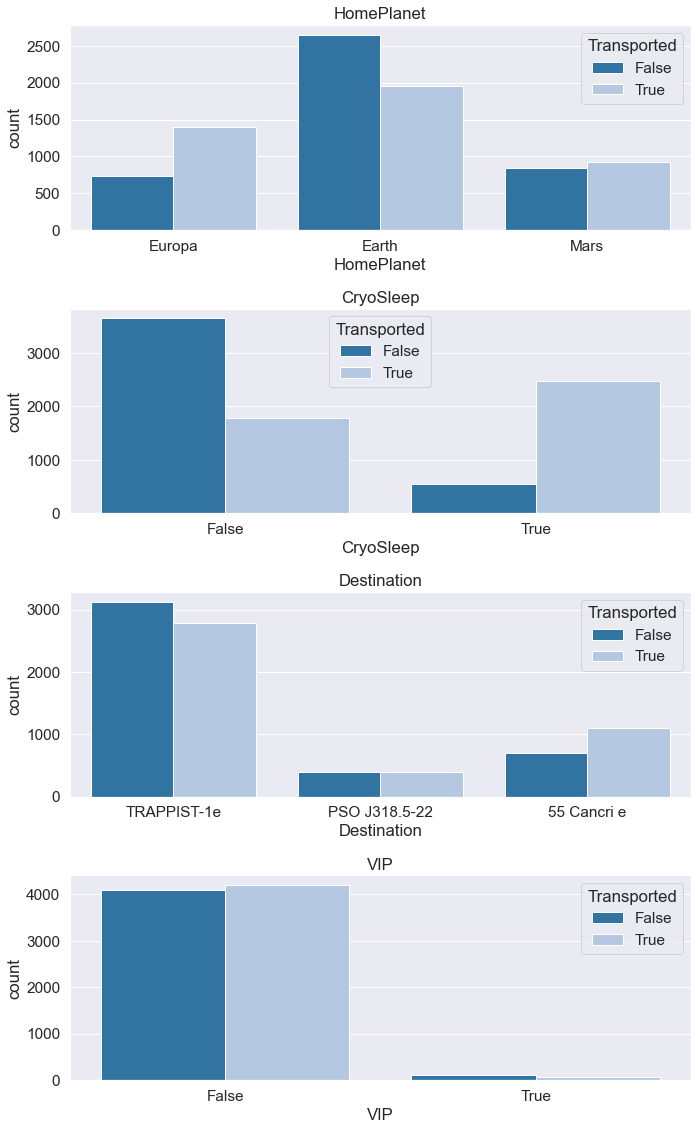


图4 各个分类计数图

从中可以得到下面的结论：

* VIP这个信息由于在传送成功和不成功的乘客里面基本都是对半分布，因此这个特征对于我们的模型训练没有什么帮助，所以我们可以将其删除
* CryoSleep很重要，因为在传送成功的乘客中，大多数都是在冷冻睡眠中醒来的

因此可以将VIP特征删除。

定性特征的分析

首先使用head()函数显示前五个记录，初步分析数据的格式。

|  |  |  |  |
| --- | --- | --- | --- |
|  | PassengerId | Cabin | Name |
| 0 | 0001\_01 | B/0/P | Maham Ofracculy |
| 1 | 0002\_01 | F/0/S | Juanna Vines |
| 2 | 0003\_01 | A/0/S | Altark Susent |
| 3 | 0003\_02 | A/0/S | Solam Susent |
| 4 | 0004\_01 | F/1/S | Willy Santantines |

表4 定性特征前5条记录

从中可以得出：

* PassengerId的形式为gggg\_pp，其中gggg表示乘客所在的组，pp表示该组中的编号。
* Cabin的形式为deck/num/side，其中side可以是P表示舷侧（Port），或者是S表示舷侧（Starboard）。

因此，

* 可以从PassengerId特征中提取组别和组内人数。
* 可以从Cabin特征中提取甲板（deck）、编号（number）和舷侧（side）。
* 可以从姓名（Name）特征中提取姓氏以识别家庭。

总之，根据上述分析，现在可以做特征工程。

特征工程

这一步是根据数据探索中获得的方向对于数据进行处理。

根据上文对年龄的分析，可以对年龄进行分组，分组结果如下：

|  |  |  |  |
| --- | --- | --- | --- |
| Age\_group | Total | Transported | Percent\_Transported |
| Age\_0-12 | 806 | 564 | 69.97519 |
| Age\_13-17 | 739 | 409 | 55.34506 |
| Age\_18-25 | 2351 | 1077 | 45.81029 |
| Age\_26-30 | 1207 | 599 | 49.62717 |
| Age\_31-50 | 2674 | 1282 | 47.94316 |
| Age\_51+ | 737 | 357 | 48.43962 |

表5 年龄分组结果

不同组内的传送的成功率有差距，基本符合本文之前数据探索当中得出的结论。

由于RoomService，FoodCourt，ShoppingMall，Spa，VRdeck这些属性都是关于各个乘客的消费数量，所以本文将它合并为一类进行分析，即消费类。

首先计算总的消费并定位没有消费的乘客。结果直方图如下所示：

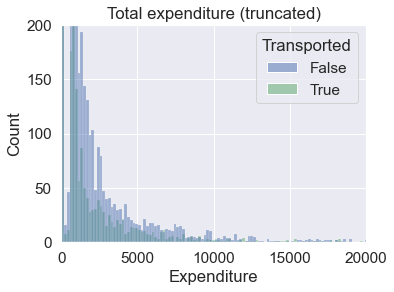


图5 消费直方图

从数据的分布可以看书，大部分乘客的消费集中在0到5000，极少部分乘客的消费超过5000。

接下来，将乘客分为消费和未消费两类，结果如下表所示。

|  |  |  |  |
| --- | --- | --- | --- |
| No\_spending | Total | Transported | Percent\_Transported |
| 0（未消费） | 5040 | 1505 | 29.86111 |
| 1（已消费） | 3653 | 2873 | 78.64769 |

表6 消费分类表

其中可以看到消费为0的乘客中，传送成功的比例要高于传送失败的比例，这可以作为一个特征。

接下来是对乘客ID进行分析。使用直方图分析如下：

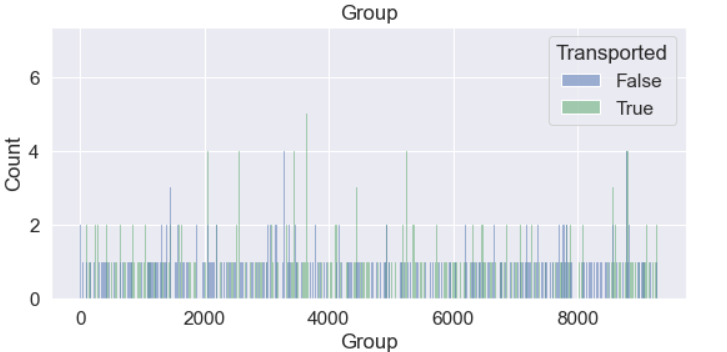


图6 乘客分组直方图

可以发现分组过多，基数太大（6217），使用one-hot会导致维度爆炸。另一方面，组的大小是一个有用的特征，因为组的人数仅有1-8这8中情况。于是根据组的大小的分类表结果如下所示：

|  |  |  |  |
| --- | --- | --- | --- |
| Group\_size | Total | Transported | Transported(100%) |
| 1 | 4805 | 2174 | 45.24454 |
| 2 | 1682 | 905 | 53.80499 |
| 3 | 1020 | 605 | 59.31373 |
| 4 | 412 | 264 | 64.07767 |
| 5 | 265 | 157 | 59.24528 |
| 6 | 174 | 107 | 61.49425 |
| 7 | 231 | 125 | 54.11255 |
| 8 | 104 | 41 | 39.42308 |

表7 分组大小表

可以发现，单独出行的人被传送的概率小于组队出行的，因此可以作为一个特征，即乘客是否单独出行。这个特征做出的直方图如下所示：

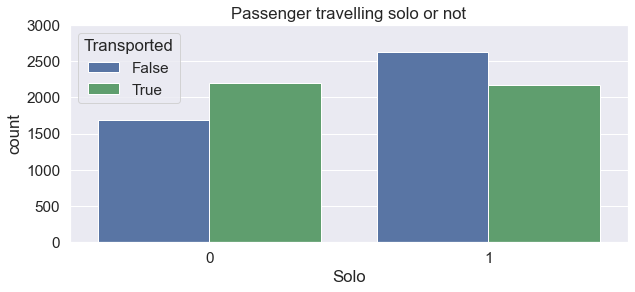
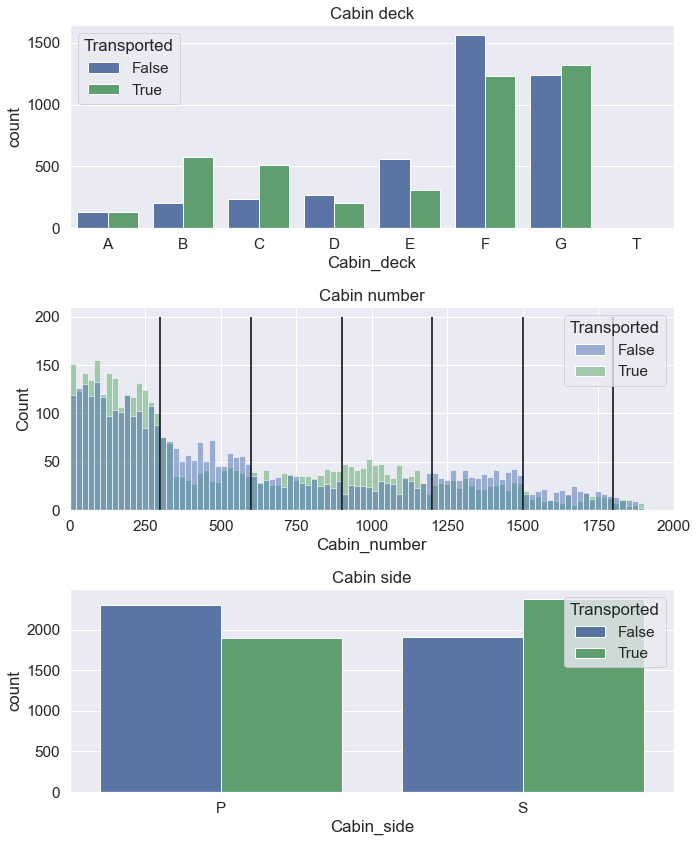
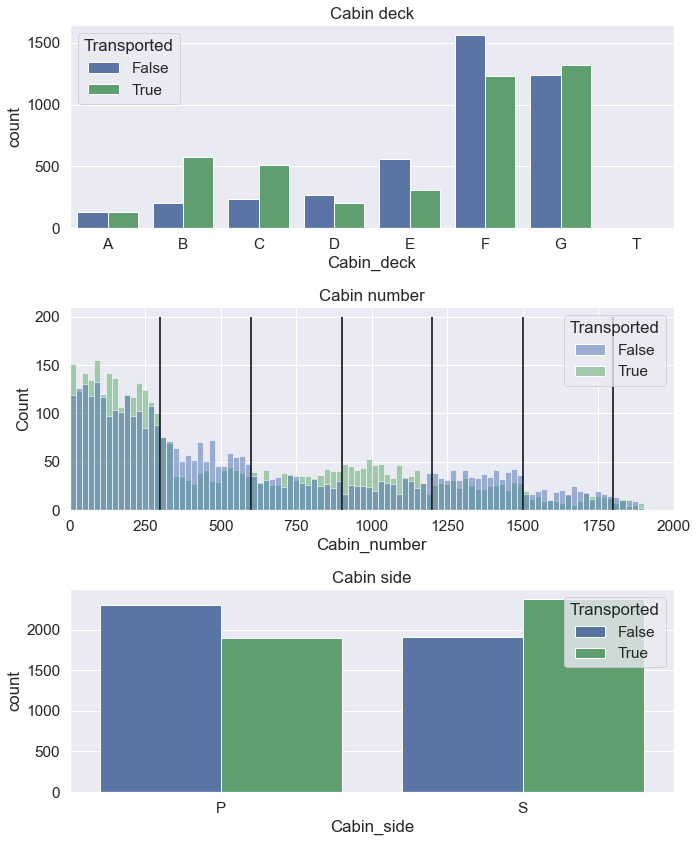


图7 乘客是否单独出行直方图

所以，乘客是否单独出行可以作为一个特征。

接下来是对Cabin信息的解析。

从上述的分析可以得知，Cabin分为三个字段，第一个字段是甲板的信息，第二个字段是编号，第三个字段是甲板的方位。分别对三个字段进行解析。做出的直方图如下所示：



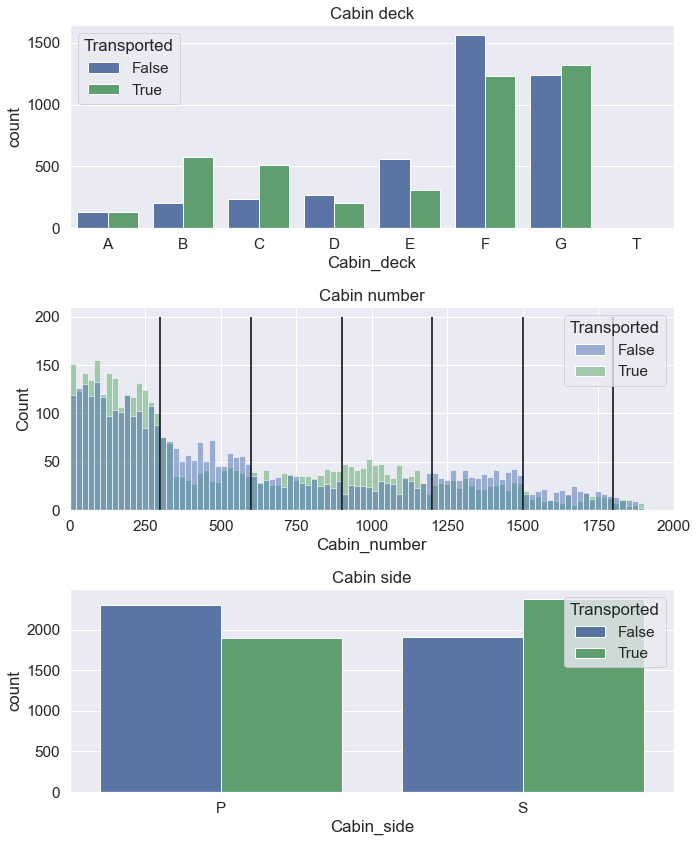


图8 Cabin三个字段直方图

对于Cabin number分组可知，被分组成了每组300个舱室。这意味着我们可以将这个特征压缩成一个分类特征，指示每位乘客所在的舱室组。

Cabin\_number分组结果如下图所示：

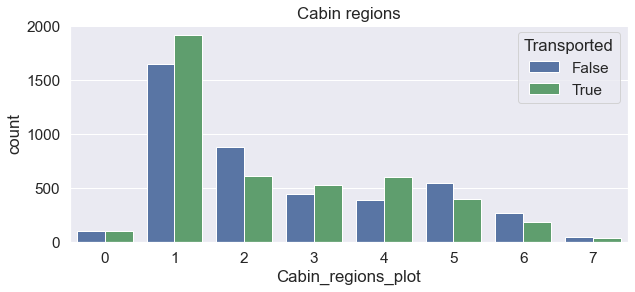


图9 Cabin\_number分组结果

因此，可以将Cabin\_number作为一个分类特征。

另外，各位乘客的姓氏可以作为一个家族的标志，根据家族也可以形成一个分类特征。

这里，对各个家族的人数进行统计，并把不同人数的家族的个数进行统计，结果如下表所示：

|  |  |  |
| --- | --- | --- |
| Family\_size | Transported | Not Transported |
| 1 | 83 | 50 |
| 2 | 228 | 197 |
| 3 | 341 | 275 |
| 4 | 425 | 380 |
| 5 | 521 | 458 |
| 6 | 535 | 474 |
| 7 | 471 | 503 |
| 8 | 357 | 396 |
| 9 | 369 | 394 |
| 10 | 251 | 332 |
| 11 | 214 | 227 |
| 12 | 168 | 176 |
| 13 | 74 | 91 |
| 14 | 100 | 108 |
| 15 | 43 | 55 |
| 16 | 42 | 52 |
| 17 | 24 | 14 |
| 18 | 21 | 22 |
| 19 | 10 | 12 |

表8 家族人数统计表

同样地，家族成员数量也可以作为一个特征。

处理缺失值

缺失值探索

在处理缺失值的时候，合并训练集和测试集来处理缺失值，两个放在一起处理缺失值更有利于从总体上把握特征，能够更好的处理缺失值。当然在完成缺失值处理之后还是要将训练集和测试集分开的。

首先，先对缺失值的分布进行探索，对于缺失值的分布和比例如下表所示：

|  |  |  |
| --- | --- | --- |
|  | Number\_missing | Percentage\_missing |
| HomePlanet | 288 | 2.22 |
| CryoSleep | 310 | 2.39 |
| Destination | 274 | 2.11 |
| Age | 270 | 2.08 |
| …… | …… | …… |

表9 缺失值分布表

由于所有的特征的缺失数量均在200-400之间，且百分比都在2%-3%之间，所以这里仅展示四个特征。接下来，使用热力图对缺失值进行进一步探索：

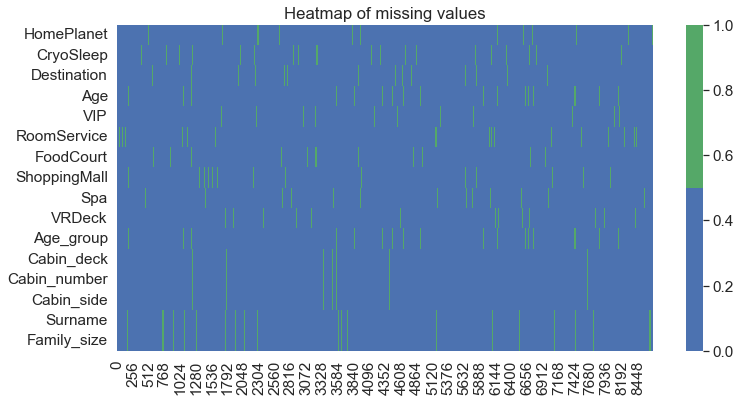


图10 缺失值热力图

从总体上看来缺失值占比大概在2%-3%，这个比例不算大。但是这可以看出，缺失值的分布并不集中，也就是说，缺失一个参数的样本数量可能会很大，所以要进一步分析。

接下来是根据缺失值的个数对样本进行统计，结果如下表所示：

|  |  |  |
| --- | --- | --- |
| Number of missing values | Count | Percentage |
| 0 | 6606 | 75.99 |
| 1 | 1400 | 16.10 |
| 2 | 421 | 4.84 |
| 3 | 217 | 2.50 |
| 4 | 40 | 0.46 |
| 5 | 6 | 0.07 |
| 6 | 3 | 0.03 |

表10 缺失值数量统计

可以看到大部分的记录缺失都是缺1个数据，有极少数的情况出现了缺了2-3个属性的，缺失3个以上属性的占比很小。

可以从中得出以下的结论：

* 缺失值与目标变量独立，大部分情况下是孤立存在的。
* 尽管数据中只有2%的缺失值，但大约25%的乘客至少有一个缺失值。
* PassengerId是唯一一个没有任何缺失值的（原始）特征。

所以，这些结论对填补缺失值有以下的帮助：

* 由于大部分缺失值是孤立存在的，与其删除行，填充这些缺失值是有意义的。
* 如果PassengerId与其他特征之间存在关联，我们可以根据该列填充缺失值。

处理缺失值的最简单方法是对连续特征使用中位数，对分类特征使用众数。这种方法足够有效，但若要最大化模型的准确性，需要寻找缺失数据中的模式。要做到这一点，可以观察特征的联合分布，例如，同一组的乘客是否倾向于来自同一家庭？显然存在许多组合，因此我们将总结我和其他人发现的有用趋势。因此，本文接下来会对这些联合分布进行研究。

联合分布分析

首先是对于HomePlanet的填补。

Group 和 HomePlanet 的联合分布如下表所示（仅展示前5组）：

|  |  |  |  |
| --- | --- | --- | --- |
| HomePlanet | Earth | Europa | Mars |
| Group |  |  |  |
| 1 | 0 | 1 | 0 |
| 2 | 1 | 0 | 0 |
| 3 | 0 | 2 | 0 |
| 4 | 1 | 0 | 0 |
| 5 | 1 | 0 | 0 |

表11 联合分布表结果示例

接着，对每一组的不同星球个数进行计数做统计，发现每一组都是来自于同一个星球，没有例外。所以，可以按照组别来补充其中来自星球的信息。

经过填补之后，HomePlanet属性的缺失值从288降低至157。

HomePlanet和CabinDeck的联合分布如下热力图所示：

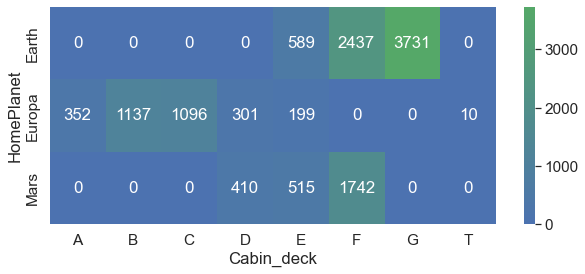


图11 HomePlanet和CabinDeck的联合分布热力图

可以得出以下结论：

* 来自Europa的乘客在 A、B、C 或 T 层甲板上。
* 来自Earth的乘客在 G 层甲板上。
* 来自多个行星的乘客在 D、E 或 F 层甲板上。

利用这样的一个规律可以将A,B,C,T,G甲板上母星缺失的给填补上。

经过填补之后，HomePlanet的缺失值从157降低至94。

HomePlanet和Surname也有与Group和HomePlanet类似的关系，即同一个姓氏的人都来自于同一个星球，因此我们可以借此规律进行填补。填补后，HomePlanet的缺失值从94降低至10。

对于剩下的10个缺失值，经画表发现这10个缺失值的目的地都是TRAPPIST-1e，所以这里研究HomePlanet和Destination之间的关系，绘制热力图如下所示：

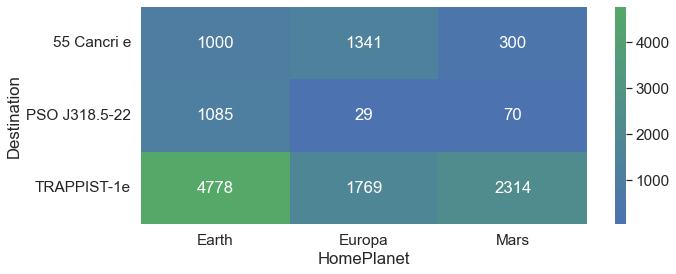


图12 HomePlanet和Destination的联合分布热力图

大多数前往TRAPPIST-1e的人来自地球，因此用Earth进行填补是合理的。没有人来自地球是在D层甲板上的，所以我们需要将它们过滤掉。因此最后将这10个缺失值填补完毕。至此，所有的HomePlanet的缺失值已经全部填补完毕。

接着是对Destination的填补。

对于上一个热力图分析可得，大部分的乘客的Destination均为TRAPPIST-1e，所以这里采用众数填补法，将Destination的缺失值用TRAPPIST-1e填补。

再者是对Surname的填补。

探索Surname和Group的关系，绘制直方图如下所示：

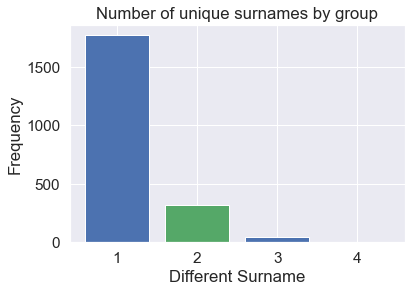


图13 Surname和Group关系直方图

大多数（83%）的群组只包含一个家庭。因此，可以根据该群组中的大多数姓氏来填充缺失的姓氏。

填补后，Surname的缺失值从294降低至155。不必去除所有这些缺失值，因为最终会删除姓氏特征。然而，本文可以据此更新家庭大小特征。

接着是对Cabin\_side的填补。探索Cabin\_side和Group之间的关系，采用与探索Surname和Group的关系相同的方法，分析可得所有的相同组的人都在甲板相同的side，因此可以利用这样一点来补充所在甲板的位置。填补之后，Cabin\_side的缺失值从299降低至162。再次探索Cabin\_side和surname之间的关系，画出，绘制出家庭成员在同一侧的甲板占比的直方图如下所示：

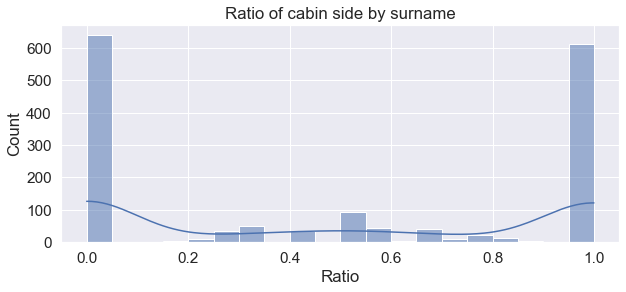


图14 家庭成员在同一侧甲板占比直方图

分析可得，家庭成员在同一侧所占的百分比 76.7 %，所以可以使用这一规律来填补缺失值。填补之后，Cabin\_side的缺失值从162降低至66。由于Cabin\_side是一个平衡的数值，剩余的数据无法草率的用其中一种去填补，所以这里全部采用异常值去填充。至此，所有的Cabin\_side的缺失值填补完毕。

接着去填补Cabin\_Deck的缺失值。首先有个前面分析出来的规律可以首先对缺失值进行初步填补，即来自相同家庭的成一半在同一个分组当中，利用这样的一个规律先来进行填补一下。填补之后，缺失值数量从299降低至162。接着进行联合分布分析，得到如下表所示的结果：

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| HomePlanet | Destination | Solo | A | B | C | D | E | F | G | T |
| Earth | 55 Cancri e | 0 | 0 | 0 | 0 | 0 | 20 | 90 | 272 | 0 |
| 1 | 0 | 0 | 0 | 0 | 47 | 289 | 269 | 0 |
| PSO J318.5-22 | 0 | 0 | 0 | 0 | 0 | 18 | 67 | 230 | 0 |
| 1 | 0 | 0 | 0 | 0 | 25 | 262 | 466 | 0 |
| TRAPPIST-1e | 0 | 0 | 0 | 0 | 0 | 133 | 438 | 1075 | 0 |
| 1 | 0 | 0 | 0 | 0 | 358 | 1350 | 1509 | 0 |
| Europa | 55 Cancri e | 0 | 96 | 377 | 313 | 59 | 35 | 0 | 0 | 2 |
| 1 | 67 | 141 | 159 | 46 | 34 | 0 | 0 | 0 |
| PSO J318.5-22 | 0 | 2 | 5 | 11 | 0 | 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 10 | 0 | 0 | 0 | 0 | 0 |
| TRAPPIST-1e | 0 | 152 | 459 | 428 | 120 | 53 | 0 | 0 | 1 |
| 1 | 44 | 179 | 201 | 84 | 82 | 0 | 0 | 8 |
| Mars | 55 Cancri e | 0 | 0 | 0 | 0 | 32 | 15 | 104 | 0 | 0 |
| 1 | 0 | 0 | 0 | 40 | 16 | 92 | 0 | 0 |
| PSO J318.5-22 | 0 | 0 | 0 | 0 | 8 | 9 | 14 | 0 | 0 |
| 1 | 0 | 0 | 0 | 9 | 7 | 21 | 0 | 0 |
| TRAPPIST-1e | 0 | 0 | 0 | 0 | 168 | 219 | 798 | 0 | 0 |
| 1 | 0 | 0 | 0 | 164 | 263 | 743 | 0 | 0 |

表12 HomePlanet、Destination、solo联合分布表

分析上述数据可以得出如下结论：

* Mars的乘客最有可能在 F 层甲板上。
* Europa的乘客（大致上）如果是独自旅行，则最有可能在 C 层甲板上，否则可能在 B 层甲板上。
* Earth的乘客（大致上）最有可能在 G 层甲板上。

利用上述规律进行填补后，缺失值数量从162降低至了0。至此Cabin\_Side填补完毕。

接着对Cabin\_number进行填补，分析Cabin\_number和Deck之间的关系，绘制出以下的散点图：

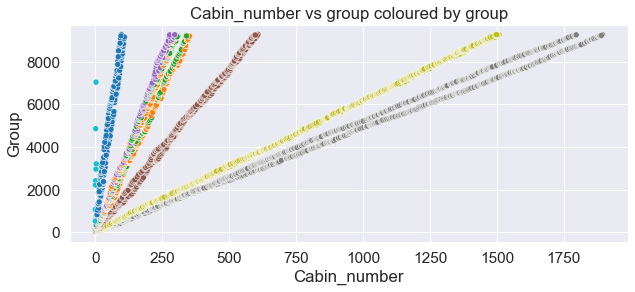


图15 Cabin\_number和Deck关系散点图

舱位号和组号在每个甲板上共享线性关系。因此，这里可以通过在每个甲板上使用线性回归来推断缺失的舱位号，以获得一个近似的舱位号。填补后，Cabin\_number的缺失值从299降低至0。

这里可以更新以下甲板分类特征工程的结果，即将填充好的Cabin\_number分配到特征工程所做的以300为阶的分类结果中。

接着是对VIP进行填充。由于VIP是一个极其不平衡的量，大部分人都是非VIP，所以采用众数填补法，即非VIP填充缺失值。

接着是对于年龄的填充。年龄在家乡星球、团队规模、花费和舱位甲板等许多特征上都存在变化，因此将根据这些子组的中位数来填补缺失值。在此，再将新填充的缺失值加入之前特征工程中的分组。

对于CryoSleep这个属性，如果一个乘客没有消费记录那他大概率会在冬眠，如果有消费记录，那在冬眠的可能性其实就不大。根据这个规律填充这个属性是合理的。

最后是对Expenditure进行填补。首先根据上述规律，即冬眠的人没有消费这个规律，将冬眠的样本用0进行填充，缺失值从1410降低至866。对于剩余的866个缺失值，本文采用以下的策略：花费在许多特征上存在差异，但只会使用家乡星球、独自旅行以及年龄组来填补缺失值，以防止过拟合。本文还将使用平均值而不是中位数，因为很大一部分乘客没有花费，中位数通常为0。并且，年龄在12岁以下的人不会有任何花费。联合分布的表格如下所示：

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| HomePlanet | Solo | Age\_0-12 | Age\_13-17 | Age\_18-25 | Age\_26-30 | Age\_31-50 | Age\_51+ |
| Earth | 0 | 0 | 724.9022222 | 789.7005545 | 841.0935961 | 736.6557734 | 733.6495726 |
| 1 | 0 | 693.0148976 | 779.3959417 | 795.4206897 | 794.8186275 | 826.3669725 |
| Europa | 0 | 0 | 1153.160256 | 2652.013298 | 3534.668246 | 3975.774005 | 3483.639004 |
| 1 | 0 | 0 | 2489.888889 | 3806 | 3949.939929 | 3952.085526 |
| Mars | 0 | 0 | 1176.839286 | 1161.808333 | 1247.098361 | 1143.671916 | 1345.419643 |
| 1 | 0 | 1687.261538 | 1075.341146 | 1107.122677 | 1110.392045 | 1100.298387 |

表13 HomePlanet、Solo、Age联合分布表

采用平均数填补后，缺失值数量降为到0。至此，所有缺失值填补完毕。

数据预处理

对于训练集和测试集，本文做了以下的预处理：

删除'PassengerId', 'Group'两个属性，因为这两个数据在特征工程中，分别用于结对出行人的个数。

删除Cabin\_number，因为此用于定性甲板的分组，这里已没有作用。

删除VIP，因为他对结果几乎没有影响。

对于消费类属性（即RoomService, FoodCourt, ShoppingMall. Spa），他们的分布差距比较大，大部分样本集中在了少部分的区域中。本文采取对数变换的方法，减小分布的偏斜程度，尤其是这里还有特别大的异常值。

现在作出分布直方图如下：（这里仅展示RoomService和FoodCount，其他规律基本相同）

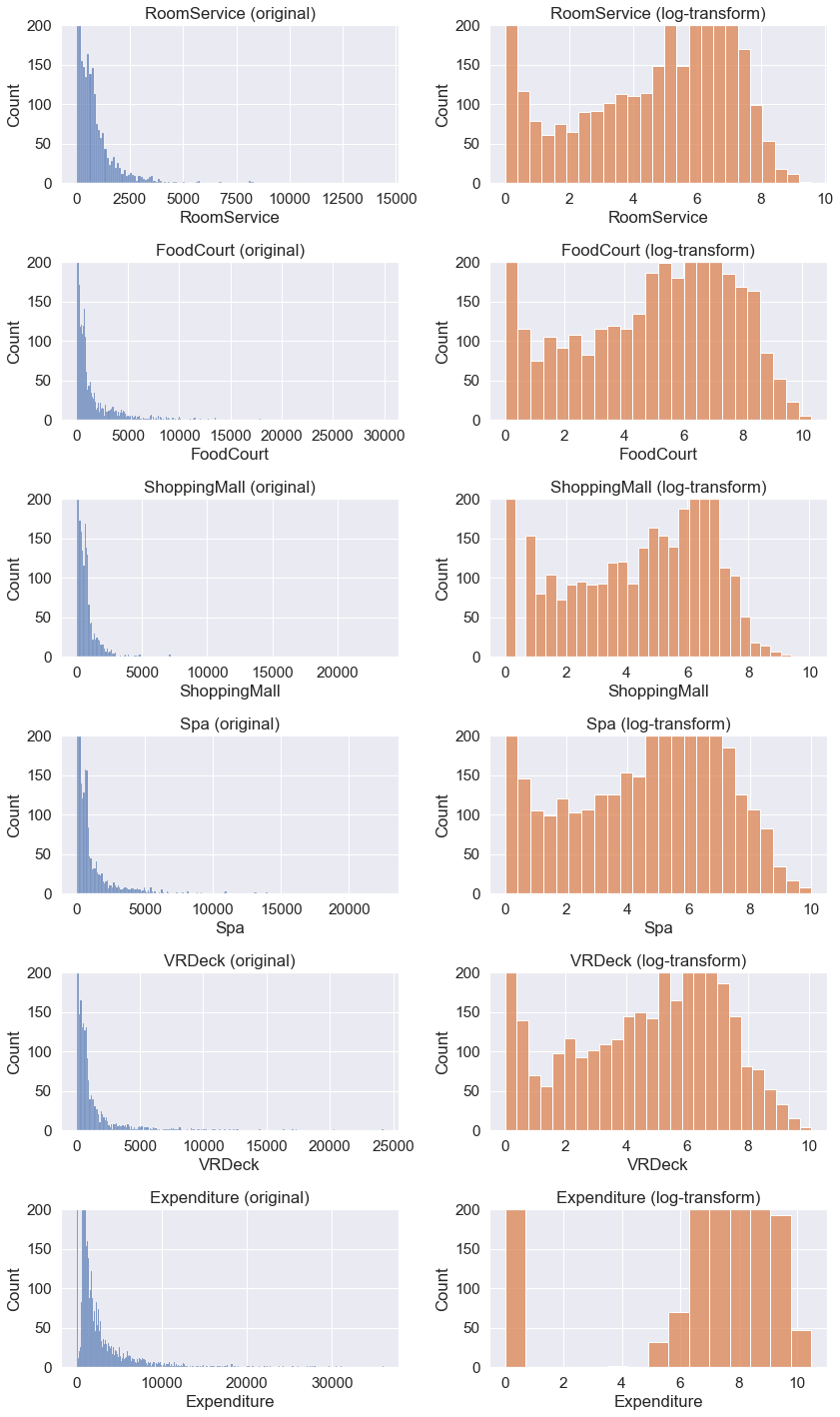


图16 取对数后的数据分布直方图

可以看到经过对数变换之后分布明显变得比之前更加均匀。

接着对数据进行编码和合理的缩放处理。步骤如下：

* 识别数值列和分类列；
* 将数值数据缩放为均值 = 0 和方差 = 1；
* One-hot编码分类数据；
* 数值预处理，分类预处理，组合预处理（使用ColumnTransformer）

经过这样的处理，可以将这些数据放入分类器进行训练。

主成分分析

使用PCA类对输入数据X进行拟合，获取每个主成分的解释方差比例。

使用matplotlib库绘制折线图，横轴为主成分数量，纵轴为解释方差的累积比例。

绘制一条红色的水平线，表示解释方差累积比例达到100%的阈值。

方差累计曲线图如下（图17）所示：

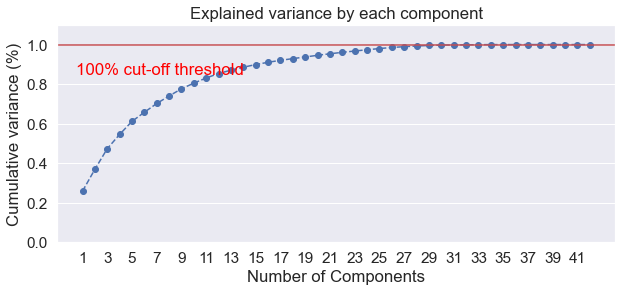


图17 方差累计曲线图

对于方差累计曲线图分析可得，在1-20曲线处于上升阶段，表示前20个主成分包含了相对重要的信息。而20以后趋于平缓，说明更多的主成分对于方差的贡献较小，重要性相对更低。这对于以后分类器的参数设置有一定的参考价值。

总结