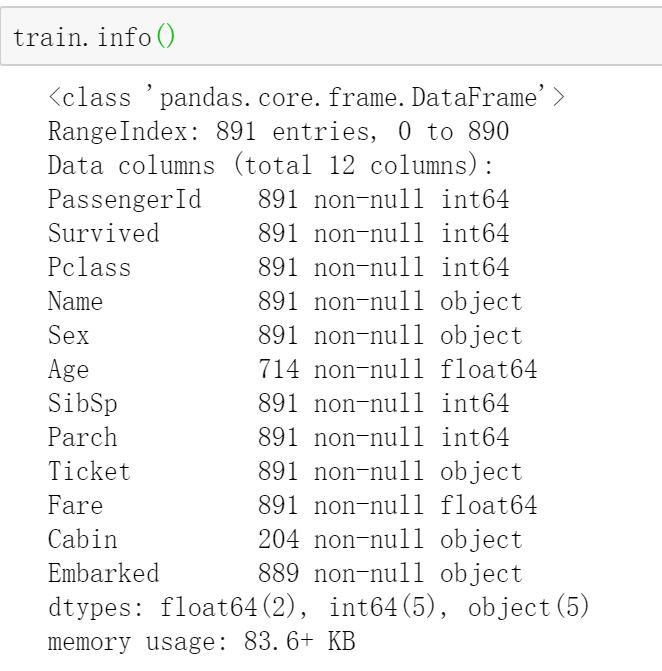
# 泰坦尼克之灾数据分析实验报告

**1707 20175364 马杰生**

## 查看特征，初步分析、猜想特征对结果的影响

****

|  |  |  |
| --- | --- | --- |
| **特征/标签名** | **初步分析，猜想** | **是否有缺失值** |
| **PassengerId(乘客ID)** | **标记乘客，对分析无用** | **/** |
| **Survived(是否生存)** | **标签，目标为二分类：生存与死亡** | **/** |
| **Pclass(船舱等级)** | **根据常识猜想，船舱越高级，越近甲板，逃生可能较快** | **/** |
| **Name(姓名)** | **姓名中含有称呼信息/头衔，**  **可以考虑提取作为有用的信息** | **/** |
| **Sex(性别)** | **妇孺在逃生时是会重点照顾，**  **理论上女性生还几率应该会比较高。** | **/** |
| **Age(年龄)** | **儿童生存率一定较高。长者可能不易生还，不排除老人有优先获救的机会，**  **目前不易结论。** | **近200个** |
| **SibSp(姐妹和配偶数)，**  **Parch(父母孩子数)** | **该乘客在船上的姐妹和配偶数量**  **该乘客在船上的父母和孩子数量**  **二者相加可看作家庭成员，家庭可能作为一个整体而获救/死亡** | **/** |
| **Ticket(票号)** | **影响未知，可能反映了人际关系（同行者）** | **/** |
| **Fare(船票价格)** | **由常识猜想，票价越高，生还机会越高** | **1个** |
| **Cabin(客舱号)** | **可能反映了船舱位置，决定了逃生的路线长短，可能有用。** | **700多个** |
| **Embarked(登船地点)** | **影响未知** | **2个** |

## 数据可视化分析

|  |  |
| --- | --- |
| **女性生存率远高于男性** |  |
| **舱位等级越高，存活可能越大** |  |
| **家庭影响：家庭（兄弟姐妹+父母子女）规模适中者，更可能存活** |  |
| **年龄影响：未成年人更容易存活（密度图橙色突出部），青壮年更可能死亡，老年人不存在生存优势（密度图蓝色突出部）** |  |
| **船票价格：价高者更容易存活** |  |

**在此，我们基本上完成了对数值型特征的可视化，余下的字符串型特征仍有相当大的使用价值，因此，我们尝试进行特征工程，将余下特征转化为可量化/更易衡量的特征。**

## 特征工程

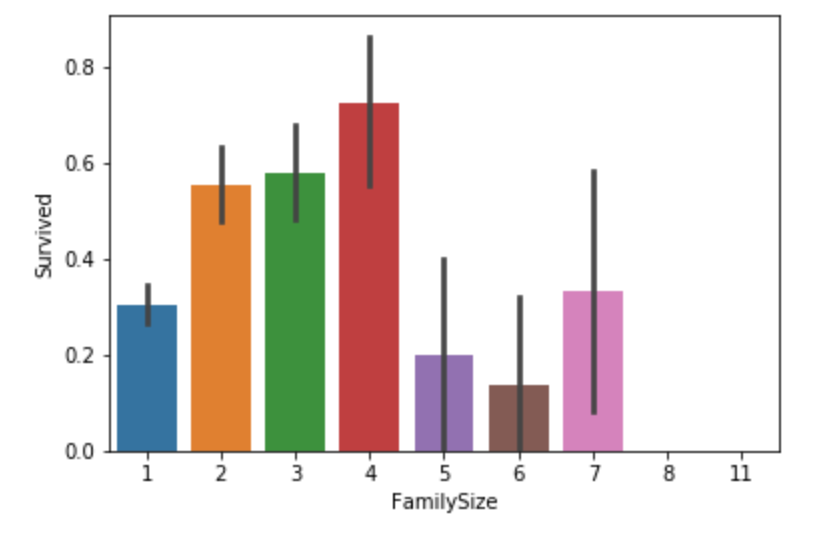
1. **姓名中头衔信息的转换：根据维基百科的资料，现在将不同的头衔/称呼进行如下的映射：**

|  |  |  |
| --- | --- | --- |
| **头衔** | **分类** | **维基百科上对于贵族的分类** |
| **'Capt', 'Col', 'Major', 'Dr', 'Rev'** | **Officer（官员/公仆等“中间阶层”）** |  |
| **'Don', 'Sir', 'the Countess', 'Dona', 'Lady'** | **Royalty（较高级的贵族）** |
| **'Mme', 'Ms', 'Mrs'** | **Mrs（普通女性）** |
| **'Mlle', 'Miss'** | **Miss（较年轻普通女性）** |
| **'Mr'** | **Mr(普通男性)** |
| **'Master','Jonkheer'** | **Master(低级贵族)** |

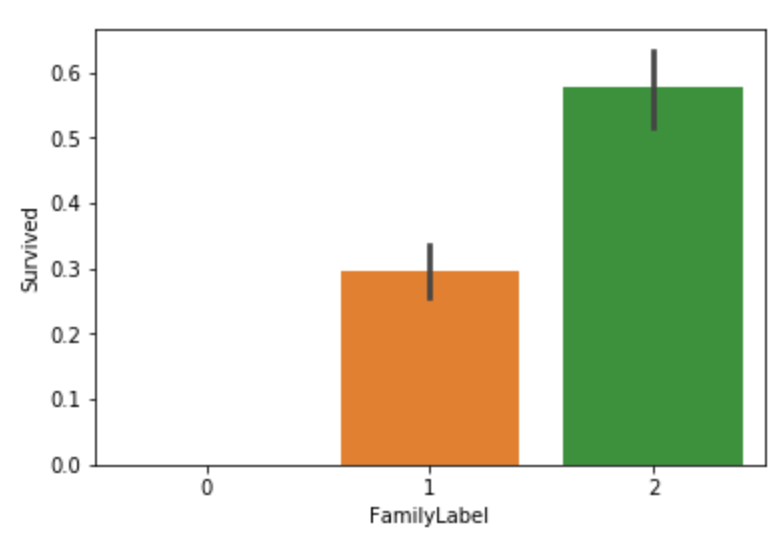
**可视化分析如下：**

|  |  |
| --- | --- |
| * **普通男性(Mr)的存活率较低；** * **女性不论身份，存活率都较高；** * **中间阶层不高，如：Capt——军官**   **Dr——学者等等；**   * **贵族（较低等的Master，高等的Royalty）的存活率也比较高** |  |

1. **将兄弟姐妹与父母配偶合并为家庭人数**

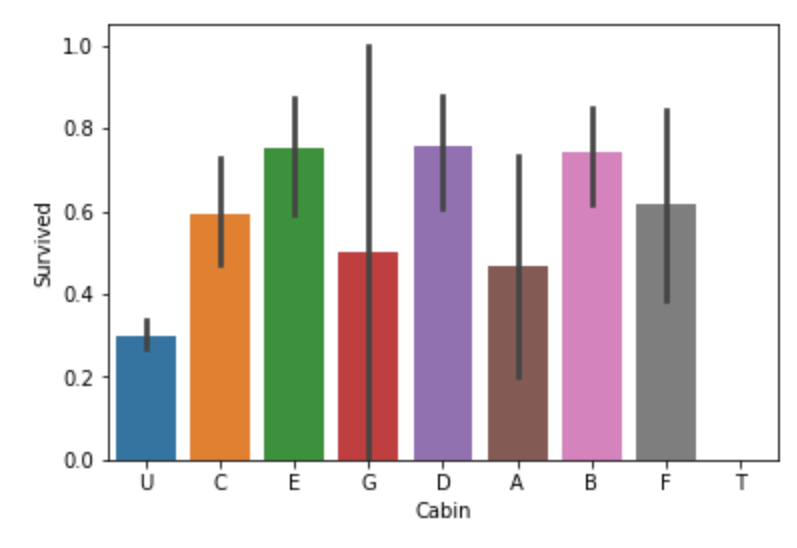
****

**7种值的特征不容易进行训练，尝试分为三类：将存活率过半的【2】、未过半的【1】、过低的【0】。其中【2】包含家庭规模为2-4的，【0】为8，11的；其余归为【1】，可视化结果如下：**

****

1. **参照以上思想，转化船票编号：**

|  |  |
| --- | --- |
|  |  |

1. **转化Cabin特征中的值，使用首字母，可视化如下**
2. **数据清洗：缺失值处理：**

**由最开头的表，可知道有四个特征存在缺失：**

|  |  |
| --- | --- |
| **Age(年龄)** | **存在缺失**  **（近200个）** |
| **Fare(船票价格)** | **存在极少缺失（1个）** |
| **Cabin(客舱号)** | **存在较多缺失（700多个）** |
| **Embarked(登船地点)** | **存在极少缺失（2个）** |

**其中：**

1. **年龄特征在之前的分析中发现较为重要，故不考虑删除，但缺失值较多不宜使用简单方法（如平均数/中位数等）处理（会增加噪声，反而降低未来模型（决策树等）预测的准确度）。所以考虑使用随机森林模型进行填充。其优点主要有以下几点:**
   * 1. **随机森林填补通过构造多棵决策树对缺失值进行填补，使得填补得到的数据具有随机性和不确定性，更能反映出这些未知数据的真实分布；**
     2. **随机森林填补由于在构造决策树过程中，每个分支节点选用随机的部分特征而不是全部特征，所以能很好的应用到高维数据的填补；**
     3. **随机森林算法本身就具有很好的分类精度，从而也更进一步确保了得到的填补值的准确性和可靠性。**

**根据可见的影响最大的四个特征：年龄、舱等、性别、头衔，构建随机森林模型预测，填补年龄缺失值。**

1. **Embarked（登船地点）特征缺失量为2，缺失Embarked信息的乘客票号相同，舱等均为1，票价均为80，通过观察数据集，这种乘客的Embarked基本上都在C，所以缺失值按此填充为C。**
2. **票价缺失量为1，缺失信息的乘客的Embarked为S，舱等为3，所以用Embarked为S，Pclass为3的乘客的Fare中位数填充。**
3. **客舱缺失值使用U代表，目前未能发掘出其对生还率的影响，先搁置处理。**
4. **注意到，一个家庭的成员大多一起存活/死亡；那些一同死亡的儿童、女性，存活的男性会对之后的模型分类造成干扰，为了使之后的分类更为准确，尝试对测试集中位于死亡家庭的妇女、儿童进行惩罚性修改，将其改为成年男性。位于幸存家庭的男性将其改为未成年女性。**
5. **最后，选择必要的特征。同时，作为分类任务，需要对数据集的非数值特征进行哑变量转化。使用get\_dummies（）方法即可。**

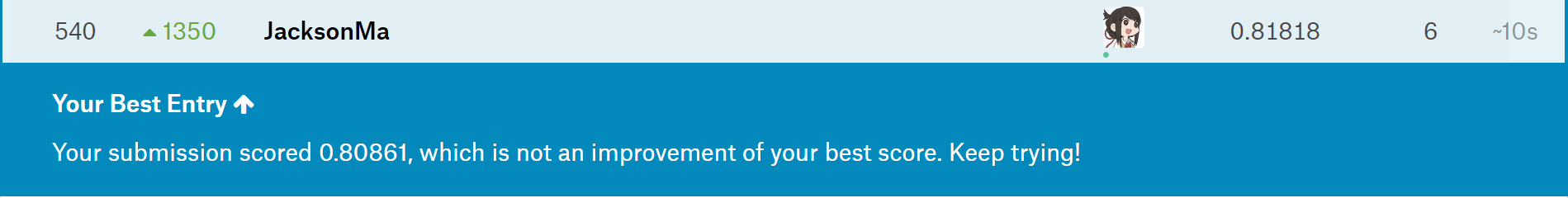
## 模型构建：

1. **根据填补缺失值的经验，与对随机森林模型的了解，本人决定继续使用随机森林作为预测模型。更为详细地阐述：因为它的基本单元决策树可以提炼出数据中的非线性趋势，因此可用于分类和回归。但决策树对训练数据的过拟合将会是一个问题，由于通过分叉，会创建非常“深”的决策树，最终得到的规律受训练数据中噪音的影响，不能一般化地推广到新数据集，换而言之，会造成过拟合。随机森林算法应运而生，随机森林包括成百上千的决策树，这些树的输入数据和分叉依据都经过一定的随机化处理。每个决策树对应的都是训练集的随机一部分，每次分叉也是基于某列的一部分数据。最后对所有树取平均，把过拟合的风险降至最低。**
2. **为提高随机森林预测精度，虽简单的办法就是增加决策树的数量。调整min\_samples\_split和min\_samples\_leaf两个参数也可以减少过拟合。因为数量不足的决策树会把训练集的噪音也拟合进模型中，从而生成过深的分支，而这不能代表真实的数据特征。所以提高min\_samples\_split和min\_samples\_leaf有助于让算法适应新数据，但是在训练集上的得分会降低。从以上分析，可知不能够贸然调大参数，否则预测结果反而更差。因此需要以有效的方法调整参数**
3. **使用网格搜索法调参：**
   1. **机器学习算法中有两类参数：从训练集中学习到的参数，比如逻辑回归中的权重参数，另一类是模型的超参数，也就是需要人工设定的参数，比如正则项系数或者决策树的深度。**
   2. **在此，本人使用了网格搜索法，通过穷举，搜寻出最适合的参数组合。经过实验提交检验，预测准确率从78%提升至将近80%，效果确实极佳。**
   3. **从网络上的经验交流，获悉对本问题而言，目前的最优参数为26棵树，深度为6，基于此再次调参。**
4. **使用Scikit-learn提供选择特征用的函数SelectKBest，自动从数据中提取最佳特征。**
5. **最终，使用pineline流水线建构模型，获得结果。交叉验证评分如下:**



## 最终结果

**排名540/11000，约前5%，之后尝试调整了几个超参数，仍未能超越**



## 回顾与反思：

1. **不足：**
   1. **对于部分特征：票号、登船地，船舱类型等未能找出更有效的利用方法。**
   2. **对于可能误导模型预测的部分样本（一等舱男性、三等舱女性等）未能找出有效的方法进行处理。**
2. **可能提升预测准确率的关键点：**
   1. **有效地融合多个模型，如逻辑回归、梯度提升等**
   2. **更深入地进行特征工程，彻底地利用之前未能有效使用的特征,对可能造成误导的样本进行处理。**