# **泰坦尼克号——来自灾难的机器学习**

李勇志、李思维、张影、崔傲叠

武汉纺织大学

武汉市江夏区阳光大道1号 邮政编码：430200

**摘要**

*本研究通过具体的实例“泰坦尼克号乘客遇难预测”。虽然幸存中涉及到一些运气成分，但似乎有些人比其他人更有可能幸存例如妇女、儿童和上层阶级。本次研究使用乘客数据（即姓名、年龄、性别、社会经济阶层等），用逻辑回归、KNN、随机森林来建立预测模型来回答以下问题：**“什么样的人更有可能幸存？”*

# **问题分析**

要回答“什么样的人更有可能幸存？”这个问题，我们需要了解问题的背景，根据背景精炼数据，选择模型，然后制定实施流程为模型进行数据准备，训练，评价，从而使用模型回答该问题。

## **问题背景**

泰坦尼克号的沉没是历史上最臭名昭著的沉船事件之一。虽然幸存下来的人存在一定的运气因素，。在灾难后，但一些人似乎更容易幸存，比如妇女，儿童和上层阶级。我们获取了乘客数据（即姓名、年龄、性别、社会经济阶层等）来预测该乘客是否会在这次灾难中存活下来。

## **模型选择**

乘客数据有12维，我们需要在其中筛选出合适的数据，筛选后大概率任为多维。我们要解决的问题就是根据所给的数据选择合适的模型，来预测测试数据集的 Survived 特征的那一列。在判断是否存活的问题上，我们可以想到的模型有：逻辑回归，XGBoost，SVM，决策树，随机森林，KNN，高斯朴素贝叶斯，voting Classifer将多个模型。最终选择逻辑回归、KNN和随机森林三种。

## **实施流程**

（1）获取训练集和测试集；

（2）进行数据整理和清洗；

（3）分析识别模式并探索数据，进行特征处理；

（4）建模、预测和解决问题。

# **数据准备**

本章分析训练和测试数据，进行数据整理，分析异常数据，补充缺失数据，并分析和Survived更相关的特征数据。

## **数据概览**

有两个相似的数据集，一个是训练集`train.csv`，另一个是测试集`test.csv`。`train.csv `包含乘客子集的详细信息（准确地说是 891 人），揭示了他们是否幸存，也称为“基本事实”。`test.csv` 数据集包含类似的信息，但没有透露每位乘客是否幸存，预测这些结果是我们的工作，即：使用你在 `train.csv` 数据中找到的模式，预测船上的其他 418 名乘客（在 `test.csv `中找到）是否幸存，数据入图1所示。



图1：训练集数据概览，

## **数据分析**

在实验中首先我们对训练集和测试集读取生成的DataFrame格式数据通过isnull方法找出缺失值并通过sum方法统计得出：

(1)、客舱号Cabin数据在训练集和测试集中都存在大量缺失，分别缺失687（共891）和327（共417）个；

(2)、年龄Age存在少量缺失，在训练集和测试集中分别缺失177（共891）和86（共417）；

接下来通过DataFrame的corr生成了相关性矩阵，观察到票价等级Pclass、乘客性别Sex与幸存与否Survived 之间有显著的相关性，因此在模型中应包含此特征。

在船的兄弟姐妹或配偶数量SibSp和在船父母或孩子数量Parch中的某些值与Survived具有相关性，最好从这些单独的特征中派生一个特征或一组特征，使得与Survived有显著相关性。

接下来我们对进行了可视化分析：

### 年龄特征分析

首先了解数值特征Age与我们的解决方案目标（Survived）之间的相关性。直方图对于分析连续数值变量很有用。

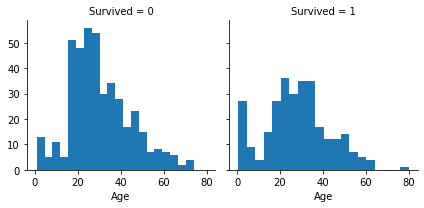


图2：罹难和幸存数据中的年龄分布

分析得到以下结论：

(1)、婴儿（年龄<=4）的存活率很高；

(2)、最年长的乘客（年龄 = 80）幸存下来；

(3)、大部分15-25 岁的人无法生存；

(4)、大多数乘客的年龄在 15-35 岁之间。

这个简单的分析证实了后续工作我们应该：

(1)、在模型训练中考虑年龄特征 Age；

(2)、完成年龄特征的缺失值处理；

(3)、捆绑年龄组。

### 票价等级PClass特征分析

由于年龄Age已经证实具有相关性，接下来一起研究年龄Age，票价等级PClass和幸存Survived之间的关系

。

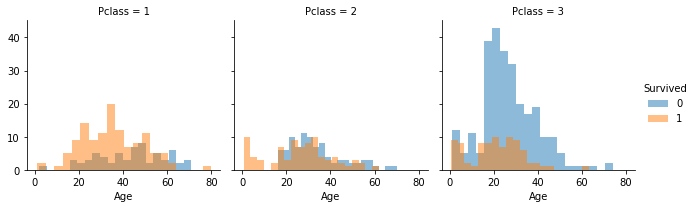


图3：年龄、票价与幸存

分析得到以下结论：

(1)、Pclass=3 有大多数乘客，但大多数没有幸存；

(2)、Pclass=2 和3 的婴儿乘客大多幸存下来；

(3)、Pclass=1 中的大多数乘客幸存下来；

(4)、Pclass 因乘客年龄分布而异

综上分析，应该将Pclass纳入模型训练之中。。

### 登船港口Embarked特征分析

分析无序分类特征 登船港口Embarked 与 Survived 相关性

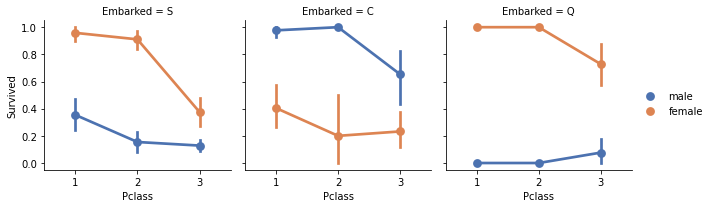


图4：登船港口、票价、性别与幸存

分析得到以下结论：

(1)、Embarked=S 和 Q 中，女性乘客的存活率远高于男性；Embarked=C 中，男性的存活率较高。这可能是 Embarked 和 Sex 相关，而 Sex 和 Survived 相关，进而造成Embarked与Survived间接相关；

(2)、对于同一票价等级和同一性别，不同登船港口的存活率不同。

综上分析，在模型训练中应该：添加 Sex 和 Embarked 特征

### 票价Fare特征分析

分析数值特征 Fare（票价） 与 Survived 相关性。

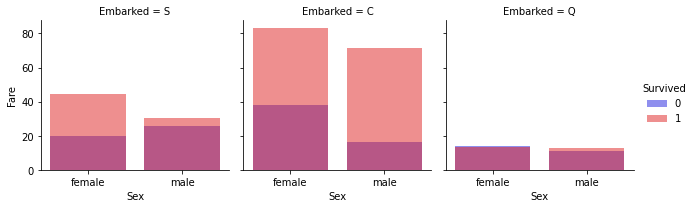


图5：登船港口、性别、费用和幸存的关系

分析得到以下结论：

(1)、支付更高票价的乘客能够更好地幸存，并且幸存率具有较明显的票价区间性；

(2)、不同的登船港口有不同的存活率。

因此，应该考虑捆绑票价特征Fare，并纳入模型训练中。

## **数据整理和清洗**

根据以上分析我们通过DataFrame的drop方法删去相关性弱切缺失较多的船舱Cabin数据，相关性较弱的票号Ticket数据，以及乘客姓名Name，乘客编号PassengerId这两个显然与幸存无否无关的数据。。

对于相关性强且没有缺失的性别Sex数据，我们将男性male替换成整形数据1，女性female替换成整形数据2便于计算。

对于相关性强且缺失的年龄Age数据，我们用平均年龄进行填补并进行了等额分箱操作讲年龄分为(-0.001, 20.0] , (20.0, 28.0] , (28.0, 29.0] , (29.0, 38.0] , (38.0, 80.0]五个区间，并赋予各个区间序数。

对于相关性强且无缺失的是否有兄弟姐妹在船上SibSp和是否有父母在船上Parch特征创建一个新特征IsAlone表示是否独自上船。。

对于相关性强少量缺失的登船港口特征Embarked，有三种可能取值 S、Q、C。仅训练数据集有两个缺失值，采用众数填补缺失值。

对相关性强且无缺失的票价Fare进行分箱并替换为序数

# **算法**

泰坦尼克号预测是机器学习中的分类问题，针对分类问题，有逻辑回归，KNN,随机森林等模型。通过运用以下这三种模型，对“什么样的人更可能幸存”进行预测。

## **逻辑回归**

### 逻辑回归模型原理

逻辑回归是机器学习中一种常见的分类模型，其对于简单的分类问题具有良好的效果。其基本原理是采用 sigmoid函数作为我们的预测函数，来预测条件概率 P（y = 1 | x）。在我们的问题中，sigmoid 函数的输出就是乘客存活下来的概率，范围在[0，1] 之间。模型在训练的过程中，通过不断最小化极大似然代价函数，来提高模型预测的准确率。在训练的过程中，加入正则化项，可在一定程度上减轻模型过拟合。

### 逻辑回归的假设函数

假设函数采用 sigmoid 函数，函数形式为如 1式，取值范围为 [0,1]。代表了每位乘客存活下来的概率。其中 z =θTxX，θ 是模型需要学习的参数，X 在该问题中对应每个乘客的特征向量。即 z 是每位乘客所有特征的线性组合。

g(z)=1/(1+e-z) (1)

### 逻辑回归的代价函数

代价函数是我们优化的目标函数，用来衡量模型在训练集上的拟合程度，在训练集上拟合的越好，代价函数就越小，在训练集上拟合的不好，那么代价函数就越大。所谓的学习过程，实质是就是不断的通过更新模型的参数值，来降低代价函数值的过程。

常见的代价函数有 MSE 代价，交叉熵代价。其中前者常常应用于回归问题中，交叉熵代价常用于分类问题中。对于我们的问题而言，由于是一个二分类问题，所以采用交叉熵代价函数。交叉熵代价函数的表达形式如 2 式。其中g(θ) 代表了逻辑回归函数的输出，log 代表以 10 为底的对数，yi 代表样本的真实分布。

J(θ)=-∑yi\*log(g(θ)) (2)

### 逻辑回归的潜在问题

理论上来说，算法在经过多次迭代的过程中。代价函数的值可以不断的降低，直到达到一个比较小的值后，基本保持不变。但是这样存在的一个问题是，模型会过度的拟合训练集的数据，从而陷入过拟合的风险。一旦模型陷入了过拟合，即使模型在训练集上的准确率很高，但是模型没有足够的泛化能力，将模型推广到未知的数据，造成在测试集上的效果不好。

### 逻辑回归的正则化

为了解决上述提到的模型可能陷入过拟合的问题，需要采取一定的措施。常见的缓解过拟合的措施，可以增加训练集的数目，或者采取正则化手段。由于我们不太容易去增加训练集的数目，因此可以采用正则化手段。常见的正则化有L1，L2 正则化。在我们的问题中采用 L1 正则化，加入正则化项的代价函数如 （3）式，其中 C 为正则化参数。[1]

J(θ)=-∑yi\*log(g(θ))+∑C|θ| (3)

## **KNN**

### KNN 模型原理

KNN（K Nearest Neighbor）算法，又称之为 K 领近算法，是数据挖掘与机器学习中最简单的分类方法之一。K领近指的是待分类样本点最近的 K 个邻居。KNN 模型最初由 Cover 和 Hart 于 1968 年提出， 是一个在理论上比较成熟的方法。

KNN 模型的主要思想是，将训练集绘制在特征空间中，然后将待分类样本，通过特定的距离计算公式，得到该样本在该特征空间最近的 K 个邻居，然后采取投票原则，将 K个邻居中得票最多的类别作为待分类样本的类别。

在我们要解决的实例问题中，我们的训练样本包括891个乘客的特征数据，将它们绘制在特征空间里。在测试集的418个数据中，我们计算每一个乘客与训练集中 891个乘客的距离远近，挑选出最近的 k 个距离，然后采取投票原则，k 个样本中所属类别最多的类别就是测试样本的类别。

### KNN模型评价

KNN的优点是对异常值不敏感，但由于我们的训练集的大小仅仅为 600 个样本， 而对于其他的一些机器学习应用，我们可能有上百万训练样本，这个时候 KNN 算法的局限性就暴露出来了，每预测一个新的样本的类别，我们都需要计算该样本与上百万样本的 距离，会造成算法运行速度非常缓慢，效率低下，这时候需要采用其他速度更快的分类模型来解决该问题。

## **随机森林**

### 随机森林原理简述

要了解随机森林，就必须要提及决策树算法。随机森林由一棵棵相同或不同的决策树组成，这个专有名词形象的表现的决策树与随机森林两者的关系。以一条数据作为一个对象，据其多项指标进行分类，每棵树都给都选择一个类别进行“投票”。在森林中票数最多者，为此数据对象的类别。

### 决策树

目前，决策树的研究主要为三种算法，分别为上世纪七十年代提出的 ID3 决策树算法，八十年代提出的 CART 决策树算法与九十年代的 C4.5 决策树算法。它们都是基于贪心算法，自上而下的构建树，其除叶子外的每个节点都选最优的属性分裂，叶子节点为此路径下样本的类别。

决策树的关键即为节点分裂时属性的抉择。属性选择度量是启发式的，它使得除叶子外的每个节点都按最优的属性分裂，在理想条件下每个划分的结果下即样本都应当为一类。

### 集成学习

因单分类器的精度不高且易出现过拟合的现象，造成分类器的泛化能力不足。集成学习则是对每个分类器所得的结果进行组合并判定样本类别，故集成学习可提升学习的泛化能力。

两个前提：其一为分类器必须是有效的，即分类器所得结果的精度高于随机；其二则为各分类器需存在差异性，使用不同的训练集或不同的学习方法即可实现。

### 随机森林

以 K 棵决策树{h（X ，θK）,k =1,2 … K}作为其基本分类器，再通过集成学习得到分类结果。因决策树培育时的样本集来自训练集的随机抽样，且此过程独立，故而{h（X ，θK）,k =1,2 … K}为独立的随机变量序列。

其决策树的培养过程如下：

从总训练集的 N 个案例中，根据bootsrap 重抽样法进行随机抽样，所得出的样本集作为这棵决策树的训练集。

若每个案例都有 M 个指标，则m<<M，m 为 M 中随机选取的指标，并据此 m 个指标计算最佳分裂方式。在整个森林的培育过程中，m 保持不变。 尽可能的培养每一棵决策树且不进行剪枝处理。

随机森林的训练即为每棵决策树的训练，因每棵树的培育过 程都相互独立，故可使用并行处理实现它的训练，如此可极大提 高决策树的培育效率。[2]

# **结论**

本次实验的结论从两方面进行分析，首先是在实验中采用的三种模型进行评价，比较三种模型的优缺点以及适应情况。然后进行幸存数据的分析，用模型模拟出的数据回答“什么样的人更有可能幸存？”这个问题。

## **模型评价**

### 逻辑回归模型评价

我们采用了sklearn 机器学习库中逻辑回归模型算法对我们的数据进行了训练，训练集生成的置信度分数为79.24，取得了良好的性能。

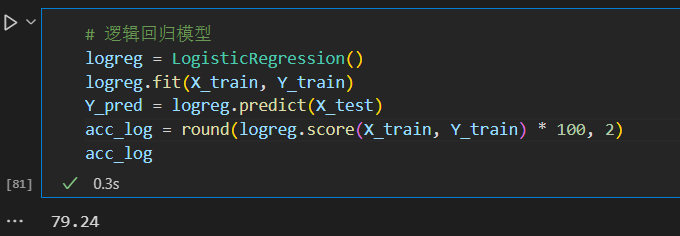


图6：逻辑回归模型评价

逻辑回归的优点是在于简单，训练速度相对于 KNN 模型快很多。但是其一般更适合用于线性可分的问题当中，而对于一些线性不可分的问题中，采用更复杂的非线性模型可能会取得更好的效果。

### KNN模型评价

我们分别测试了当 k 取 3，5，7，9 时模型的结果。测试的结果显示，在上述 4 种 k 的取值下，在测试集上的置信度分数分别为83.61%，84.18%，83.39%，83.16%。KNN的优点是对异常值不敏感，但由于我们的训练集的大小仅仅为 600 个样本， 而对于其他的一些机器学习应用，我们可能有上百万训练样本，这个时候 KNN 算法的局限性就暴露出来了，每预测一个新的样本的类别，我们都需要计算该样本与上百万样本的 距离，会造成算法运行速度非常缓慢，效率低下，这时候需要采用其他速度更快的分类模型来解决该问题。

KNN的优点是对异常值不敏感，但由于我们的训练集的大小仅仅为 600 个样本， 而对于其他的一些机器学习应用，我们可能有上百万训练样本，这个时候 KNN 算法的局限性就暴露出来了，每预测一个新的样本的类别，我们都需要计算该样本与上百万样本的 距离，会造成算法运行速度非常缓慢，效率低下，这时候需要采用其他速度更快的分类模型来解决该问题。

### 随机森林模型评价

我们采用sklearn机器学习库中随机森林模型算法对我们的数据进行了训练，训练集生成的置信度分数为86.76，证明随机森林模型算法有很好的分类作用。

随机森林相对于KNN和逻辑回归的准确率得分最高，缺点为识别不明显的分类问题上会产生过拟合行为。

## **幸存分析**

### 年龄

- 婴儿（年龄<=4）的存活率很高；

- 最年长的乘客（年龄 = 80）幸存下来；

- 大部分15-25 岁的人无法生存；

### 票价等级

- Pclass=3 有大多数乘客，但大多数没有幸存；

- Pclass=2 和 Pclass=3 的婴儿乘客大多幸存下来；

- Pclass=1 中的大多数乘客幸存下来；

### 登机港口

- Embarked=S 和 Q 中，女性乘客的存活率远高于男性；Embarked=C 中，男性的存活率较高。这可能是 Embarked 和Sex相关，而Sex 和 Survived相关，进而造成Embarked与Survived间接相关；

- 对于同一票价等级和同一性别，不同登船港口的存活率不同。

### 票价费用

- 支付更高票价的乘客能够更好地幸存，并且幸存率具有较明显的票价区间性；

- 不同的登船港口有不同的存活率。

# **引用**

1. 王可晴. 基于泰坦尼克之灾问题的机器学习传统算法和神经网络算法对比分析[J]. 电子制作,2019(2):37-39. DOI:10.3969/j.issn.1006-5059.2019.02.015.
2. 闻锋,张世奇. 浅论随机森林算法的原理与应用[J].电脑迷,2018(3):130. DOI:10.3969/j.issn.1672-528X.2018.03.123.