**Spaceship Titanic预测**

**背景：**

欢迎来到 2912 年，您需要具备数据科学技能才能解开宇宙之谜。我们收到了来自四光年之外的信号，情况看起来不太妙。

泰坦尼克号宇宙飞船是一个月前发射的星际客轮。这艘船载有近 13,000 名乘客，将移民从太阳系运送到围绕附近恒星运行的三个新宜居系外行星。

在绕过半人马座阿尔法星前往其第一个目的地——炎热的巨蟹座 55 E 的途中，粗心的宇宙飞船泰坦尼克号与隐藏在尘埃云中的时空异常相撞。可悲的是，它遭遇了与 1000 年前同名的类似命运。虽然这艘船完好无损，但几乎一半的乘客都被传送到了异次元！

为了帮助救援人员和找回失踪的乘客，您面临的挑战是使用从飞船损坏的计算机系统中恢复的记录来预测哪些乘客被异常运送。

帮助拯救他们并改变历史！

**一、数据集介绍**

* 名称：spaceship-titanic
* 属性：

|  |  |  |
| --- | --- | --- |
| Variable | Definition | Key |
| PassengerId | 乘客ID | 每个 Id 都采用 gggg\_pp 的形式，其中 gggg 表示乘客所在的团体，pp 是他们在团体中的编号。一个团体中的人通常是家庭成员，但并非总是如此. |
| HomePlanet | 乘客离开的星球，通常是他们永久居住的星球。 |  |
| CryoSleep | 表示乘客是否选择在航行期间处于假死状态。冷冻睡眠中的乘客被限制在他们的船舱内。 |  |
| Cabin | 旅客下榻的舱位号。 | 采用 deck/num/side 的形式，其中 side 可以是 P 或S |
| Destination | 目的地 |  |
| Age | 乘客年龄 |  |
| VIP | 旅客是否在航程中支付了特殊贵宾服务费用 |  |
| RoomService&FoodCourt&ShoppingMall&Spa&VRDeck | 乘客在泰坦尼克号宇宙飞船的众多豪华设施中每一项所支付的费用。 |  |
| Name | 乘客姓名 |  |
| Transported | 乘客是否被传送到另一个维度。这是目标，您要预测的列。 |  |

* 提供文件介绍：

训练数据集：train.csv ，约8700条数据。

测试数据集：test.csv，约4300条数据。

结果文件示例：sample\_submission.csv

**二、数据处理**

1、查看数据情况，如是否有缺失值和异常值，并对其进行处理。

2、分析属性与标签的关联程度。

3、特征工程，根据现有特征构建新特征，挖掘数据新的特点。

**三、模型设计**

选择以下五种模型方法进行实现，并进行比较。

1. **支持向量机：调用API实现**
2. **CART决策树：要求手写代码**
3. **以CART决策树为基学习器的AdaBoost：要求手写代码**
4. **以CART决策树为基学习器的随机森林：要求手写代码**
5. **集成学习：调用API实现，不限定基学习器和集成算法，以取得更好的性能为主要目标**

**注：**将五种不同模型封装成函数，分别命名为**Svm()、Cart()、Ada()、Rf()和MyEL()**

**四、预测结果**

根据上一步构建的模型进行预测，并将预测结果写入**result\_svm.csv**、**result\_cart.csv、result\_ada.csv、result\_rf.csv和result\_myel.csv**文件中。

**五、实验要求**

提交要求：

* 代码源文件：

**程序源码**、**预测结果文件**result\_xxx.csv和**kaggle运行后相应结果图（5张，每张图应包含模型在kaggle上的得分score）。其中：**

result\_xxx.csv文件应包含两列：

（1）PassengerId-测试集中每位乘客的 ID。

（2）Transported-对于每位乘客，预测 True 或 False

* 实验报告：描述实验方案和结果

包括但不限于：

**（1）实验目的**

**（2）数据预处理**

**（3）模型设计与选择**

**（4）预测结果**

* 将上述文件放在一个文件夹中，并命名为：学号\_姓名，上传至课程平台。

参考：<https://www.kaggle.com/competitions/spaceship-titanic/data>