

按照软件学报格式规范。

题目：

欢迎来到2912年，在这里需要你的数据科学技能来解决一个宇宙之谜。我们收到了来自四光年外的传输，情况看起来并不好。

宇宙飞船泰坦尼克号是一个月前发射的星际客轮。船上有近13，000名乘客，这艘船开始了处女航，将移民从我们的太阳系运送到三颗围绕附近恒星运行的新的可居住系外行星。

在绕半人马座阿尔法星前往其第一个目的地 - 炎热的55 Cancri E的途中，粗心的泰坦尼克号宇宙飞船与隐藏在尘埃云中的时空异常相撞。可悲的是，它遇到了与1000年前同名的命运。虽然飞船完好无损，但几乎一半的乘客被运送到另一个维度！

为了帮助救援人员并找回失踪的乘客，您面临的挑战是使用从宇宙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。

帮助保存它们并改变历史记录！

评估

提交根据其分类准确性（正确预测标签的百分比）进行评估。

提交格式

比赛的提交格式是具有以下格式的csv文件：

PassengerId,Transported

0013\_01,False

0018\_01,False

0019\_01,False

0021\_01,False

etc.

数据集描述

在这场比赛中，你的任务是预测在泰坦尼克号宇宙飞船与时空异常相撞期间，乘客是否被运送到另一个维度。为了帮助你做出这些预测，你会得到一组从船上损坏的计算机系统中恢复的个人记录。



问题：

本次课程设计将介绍关于飞船乘客的去向预测问题。这是一个竞赛任务。这个问题要求我们将已知去向的大约三分之二的乘客作为训练集，来预测未知去向的另外的三分之一的乘客。

本报告基于kaggle比赛的数据集，在本问题以科幻为背景，要求根据已有的乘客的个人信息并建立合理的模型，对于在宇宙飞船失事之后乘客的去向进行分析、预测。

数据集大约有13000个样本，12个特征参数，1个标签（待预测项）。这12个参数均会对乘客的去向产生影响。

对于这个问题，由于数据有缺失值，我们将首先寻找合适的方法对数据进行预处理，这里设想探索XGboost，LightGBM，Catboost等处理缺失值的方法。随后对特征值的分布进行分析并对特征进行编码方便后续处理。随后尝试使用10种不同的分类器，如AdaBoost、随机森林、逻辑回归等进行预测和分类，对于得到的结果进行可视化后进行比较，选出效果最好的。最终选取在之前所试验的分类器中表现性能最为优越的进行五折交叉验证并且分析各个特征对于结果的影响、分析各个属性对于结果的影响并得到预测结果。