**报告**

**学号: 姓名: 班级:**

1. **目标**

使用python完成对灾难数据集Titanic Machine Learning的全过程数据分析和预测。

1. **任务**

按照以下实验步骤完成实验报告。

Data Dictionary

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Definition | type | Explain |
| PassengerId | Id of passengers | int | Train data: 1-891  Test data: 892-1309 |
| Survived | Survived or not | bool | 0:No 1:Yes |
| Pclass | Ticket class | int | 1: 1st 2: 2nd 3:3rd |
| Name | Name of passengers | string |  |
| Sex | Sex of passengers | string | Male female |
| Age | Age in years | int |  |
| SibSp | # of siblings / spouses aboard the Titanic | int |  |
| ParCh | # of parents / children aboard the Titanic | int |  |
| Ticket | Ticket number | string |  |
| Fare | Passenger fare | string |  |
| Cabin | Cabin number | string |  |
| Embarked | embarked Port of Embarkation | string | C = Cherbourg,  Q = Queenstown,  S = Southampton |

Variable Notes

1. pclass: A proxy for socio-economic status (SES)

1st = Upper

2nd = Middle

3rd = Lower

2. age: Age is fractional if less than 1. If the age is estimated, is it in the form of xx.5

3. sibsp: The dataset defines family relations in this way...

Sibling = brother, sister, stepbrother, stepsister

4. Spouse = husband, wife (mistresses and fiancés were ignored)

5. parch: The dataset defines family relations in this way...

Parent = mother, father

Child = daughter, son, stepdaughter, stepson

Some children travelled only with a nanny, therefore parch=0 for them.

**2.1 数据清洗**

并不是每个变量都是有用的，也不是每个变量都适合建立模型，所以我们需要从原有的变量中提取信息来创建新的特征。

1. Name变量

在Name变量中，每个人都有一个Passenger头衔。根据这个标题，我们可以用它来代替Name变量。比如：我们可以将标题进行划分，将低频标题合并成一个类。

For example: 'Dona', 'Lady', 'the Countess','Capt', 'Col', 'Don', 'Dr', 'Major', 'Rev', 'Sir', 'Jonkheer' can be merged.

Mlle <- ‘Miss’

Ms <- ‘Miss’

‘Mme’<-‘Mrs’

1. Family size 变量:

我们可以在SibSp和Parch变量的基础上创建一个新的Family size变量。用柱状图列出家庭数量与存活率之间的关系。并对变量进行分化，例如大小{1,2-4,>4}或者其他.

1. More variables:

在Cabin变量中，我们可以得到甲板信息，也就是Cabin的第一个字母。

**2.2 缺省值**

方法:

1) 删除整行或整列

2）使用平均值

3）使用中位数

4) 预测

例如如下，你可以使用合适的方法来填充缺失的值：

1. missing Embarked

use Median? Or same Passenger class and fare have same Embarked?

1. Missing Fare
2. Missing Age

Prediction method

……

**2.3 特征工程**

创建‘children’ and ‘mother’ 变量

|  |  |
| --- | --- |
| Variable | Definition |
| children | 0: child :Age<18  1: adult: Age>18 |
| Mother | 0: mother: female, adult, have one or more children, title is not ‘Miss’  1: not mother |

**2.4 模型构建和预测**

1.使用训练数据建立模型

2.利用测试数据验证模型。

3.使用错误率评价指标。