**数据挖掘实验报告**

**题目 泰坦尼克获救预测**

**姓名 郑阳**

**学号 1811082065**

# 题目背景及要求

泰坦尼克号的沉没是历史上最严重的沉船事故之一。 1912年4月15日，在她的处女航中，泰坦尼克号在与冰山相撞后沉没，在2224名乘客和机组人员中造成1502人死亡。这场耸人听闻的悲剧震惊了国际社会。造成海难失事的原因之一是乘客和机组人员没有足够的救生艇。尽管幸存下沉有一些运气因素，但有些人比其他人更容易生存，比如妇女，儿童和上流社会。

要求完成对哪些人可能存活的分析。

提供两份数据，训练数据train.csv,测试数据test.csv。最终预测test.csv中乘客的获救情况。

# 做题流程

1. 引用的库

2. 读取数据集并大概分析数据

3. 处理缺失值

4. 数据可视化分析

5. 特征编码与融合

6. 特征标准化

7. 建模

8. 模型评估与选择

1. 引用库

***Numpy, pandas, seaborn, pyplot***



1. 读取数据集并大概分析数据

***training = pd.read\_csv("train.csv")***

***testing = pd.read\_csv("test.csv")***

***training.head()***

| **PassengerId** | **Survived** | **Pclass** | **Name** | **Sex** | **Age** | **SibSp** | **Parch** | **Ticket** | **Fare** | **Cabin** | **Embarked** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 1 | 0 | 3 | Braund, Mr. Owen Harris | male | 22.0 | 1 | 0 | A/5 21171 | 7.2500 | NaN | S |
| **1** | 2 | 1 | 1 | Cumings, Mrs. John Bradley (Florence Briggs Th... | female | 38.0 | 1 | 0 | PC 17599 | 71.2833 | C85 | C |
| **2** | 3 | 1 | 3 | Heikkinen, Miss. Laina | female | 26.0 | 0 | 0 | STON/O2. 3101282 | 7.9250 | NaN | S |
| **3** | 4 | 1 | 1 | Futrelle, Mrs. Jacques Heath (Lily May Peel) | female | 35.0 | 1 | 0 | 113803 | 53.1000 | C123 | S |
| **4** | 5 | 0 | 3 | Allen, Mr. William Henry | male | 35.0 | 0 | 0 | 373450 | 8.0500 | NaN | S |

**有以下这些字段**  
PassengerId => 乘客ID  
Pclass => 乘客等级(1/2/3等舱位)  
Name => 乘客姓名  
Sex => 性别  
Age => 年龄  
SibSp => 堂兄弟/妹个数  
Parch => 父母与小孩个数  
Ticket => 船票信息  
Fare => 票价  
Cabin => 客舱  
Embarked => 登船港口

2.1查看数据情况

***training.info()***

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 891 entries, 0 to 890

Data columns (total 12 columns):

PassengerId 891 non-null int64

Survived 891 non-null int64

Pclass 891 non-null int64

Name 891 non-null object

Sex 891 non-null object

Age 714 non-null float64

SibSp 891 non-null int64

Parch 891 non-null int64

Ticket 891 non-null object

Fare 891 non-null float64

Cabin 204 non-null object

Embarked 889 non-null object

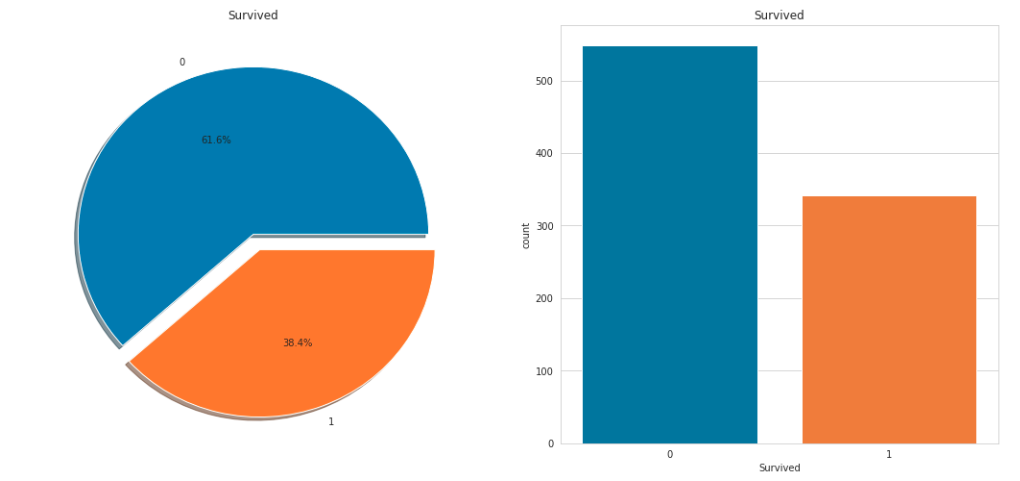
dtypes: float64(2), int64(5), object(5)

memory usage: 83.6+ KB

训练数据中总共有891名乘客，但属性的数据不全

* Age（年龄）只有714名乘客有记录
* Cabin（客舱）只有204名乘客有记录

同时发现只有只有年龄与船票价格为连续值，而性别（男，女） 登船地点（S,Q,C）为离散值。



通过画出获救图可以看出大部分人没有生还，生还者 342 人, 549 人未生还

1. 处理缺失值

对于特征缺失值的处理有4种方法。

* 方法1：直接使用均值/中位数进行填充
* 方法2：根据经验值填充
* 方法3：回归模型预测
* 方法4：剔除掉

Note:

* 1. 如果缺值的样本占总数比例极高，可以直接舍弃
  2. 如果缺值的样本适中，而该属性非连续值特征属性(比如说类目属性)，那就把NaN作为一个新类别，加到类别特征中
  3. 如果缺值的样本适中，而该属性为连续值特征属性，有时候我们会考虑给定一个step，然后把它离散化，之后把NaN作为一个type加到属性类目中。
  4. 有些情况下，缺失的值个数并不是特别多，也可以试着根据已有的值，拟合一下数据，补充上。
  5. 首先查看缺失值的特征

Training Data Missing

PassengerId 0

Survived 0

Pclass 0

Name 0

Sex 0

Age 177

SibSp 0

Parch 0

Ticket 0

Fare 0

Cabin 687

Embarked 2

dtype: int64

Testing Data Missing

PassengerId 0

Pclass 0

Name 0

Sex 0

Age 86

SibSp 0

Parch 0

Ticket 0

Fare 1

Cabin 327

Embarked 0

dtype: int64

**发现缺失的特征为Age（年龄）,Cabin（客舱号）,Embarked（登陆点）**

* + 1. **对于Cabin缺失过多，先分析统计情况**

***training.Cabin.value\_counts()***

C23 C25 C27 4

G6 4

B96 B98 4

D 3

C22 C26 3

E101 3

F2 3

F33 3

B57 B59 B63 B66 2

C68 2

B58 B60 2

E121 2

D20 2

E8 2

E44 2

B77 2

C65 2

D26 2

E24 2

E25 2

B20 2

C93 2

发现Cabin的分布过于离散，且因为缺失量过大，所以连同无关变量Ticket（船票编号）一同删掉。

* + 1. **对于Embarked，因为只缺失了2个，观察其分布，用众数进行填充**

***training.Embarked.value\_counts()***

S 644

C 168

Q 77

Name: Embarked, dtype: int64

***training['Embarked'].fillna('S', inplace=True)***

***null\_value(training, testing)***

* + 1. **对于Fare，只在测试集缺失一条，直接用中位数填充**

***testing['Fare'].fillna(testing['Fare'].median(), inplace=True)***

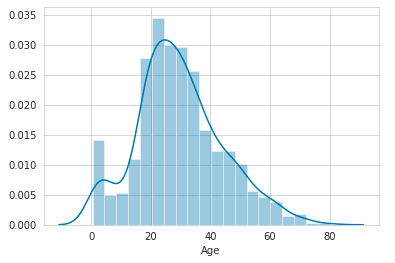
***null\_value(training, testing)***

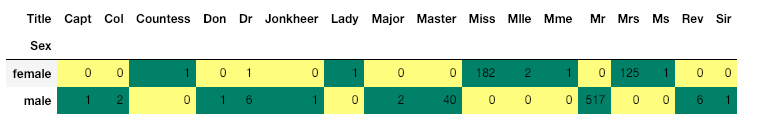
* + 1. **对于Age，首先观察其分布情况**

***copy = training.copy()***

***copy.dropna(inplace=True)***

***sns.distplot(copy.Age)***



因Age为连续值，可尝试分为多个step，来将其离散化。通过检查名字特征可以看到名字有像先生或夫人这样的称呼，这样就可以把先生和夫人的平均值分配给各自的组。

将所有的称呼替换为Master,Miss,Mr,Mrs,Other五种，再取均值对每个称呼的Age重新赋值。

***training.groupby('Title')['Age'].mean()***

Title

Master 4.574167

Miss 21.860000

Mr 32.739609

Mrs 35.981818

Other 45.888889

***def set\_missing\_age(data):***

***# 使用每组的均值来进行填充***

***data.loc[(data.Age.isnull()) & (data.Title=='Master'), 'Age'] = 5***

***data.loc[(data.Age.isnull()) & (data.Title=='Miss'), 'Age'] = 22***

***data.loc[(data.Age.isnull()) & (data.Title=='Mr'), 'Age'] = 33***

***data.loc[(data.Age.isnull()) & (data.Title=='Mrs'), 'Age'] = 36***

***data.loc[(data.Age.isnull()) & (data.Title=='Other'), 'Age'] = 46***

***set\_missing\_age(training)***

***set\_missing\_age(testing)***

最终缺失值处理完毕，所有缺失值都已填充完毕。

Training Data Missing

PassengerId 0

Survived 0

Pclass 0

Sex 0

Age 0

SibSp 0

Parch 0

Fare 0

Embarked 0

Title 0

dtype: int64

Testing Data Missing

PassengerId 0

Pclass 0

Sex 0

Age 0

SibSp 0

Parch 0

Fare 0

Embarked 0

Title 0

dtype: int64

1. 数据可视化分析
2. 特征编码与融合
3. 特征标准化
4. 建模
5. 模型评估与选择