数据集

Kaggle-波士顿房价预测数据集

数据集中包含79个特征变量，全部含义翻译如下：

- SalePrice：目标变量，表示房屋的销售价格。这是我们需要预测的变量。
- MSSubClass：房屋的建筑类型，如一层、半层、双层等。
- MSZoning：房屋所在地区的用途，如住宅区、商业区等。
- LotFrontage：与房屋相邻的街道的线性英尺数。
- LotArea：占地面积，单位是平方英尺。
- Street：房屋所在街道的类型，是砾石路还是铺设路。
- Alley：小巷通道的类型，有些房屋没有小巷通道。
- LotShape：房屋的占地面积形状，如矩形、不规则等。
- LandContour：房屋的地形，如平坦、陡峭等。
- Utilities：房屋所享有的实用工具类型，如电、气、水等。
- LotConfig：房屋在街区内的位置，如内部、角落等。
- LandSlope：房屋的地形坡度，如平原、缓坡等。
- Neighborhood：房屋所在的社区，有些社区更受欢迎。
- Condition1 和 Condition2：房屋所在的主要路径或铁路的噪音或接近类型。
- BldgType：房屋的建筑类型，如单户、双户等。
- HouseStyle：房屋的风格，如1层，1.5层，2层等。
- OverallQual：房屋的整体材料和装修质量。
- OverallCond：房屋的整体条件，如房屋年龄、维修等。
- YearBuilt：房屋建造的年份。
- YearRemodAdd：房屋的改建年份和扩建年份。
- RoofStyle：房屋的屋顶类型，如平顶、倾斜等。
- RoofMatl：房屋的屋顶材料，如瓦片、沥青等。
- Exterior1st 和 Exterior2nd：房屋的外立面材料，如木板、砖块、涂料等。
- MasVnrType：砌体饰面类型，如石头、砖块等。

- MasVnrArea：面积。
- ExterQual：外立面材料和装修质量。
- ExterCond：外立面材料的当前状态。
- Foundation：房屋的地基类型。
- BsmtQual：地下室的高度。
- BsmtCond：地下室的当前状态。
- BsmtExposure：地下室墙面的暴露程度。
- BsmtFinType1 和 BsmtFinType2：地下室装修程度。
- TotalBsmtSF：地下室的总面积。
- Heating：房屋的供暖类型。
- HeatingQC：供暖质量和条件。
- CentralAir：是否有中央空调。
- Electrical：房屋的电力系统。
- 1stFlrSF 和 2ndFlrSF：第一层和第二层的面积。
- LowQualFinSF：低品质装修的面积。
- GrLivArea：地上生活区域的面积。
- BsmtFullBath 和 BsmtHalfBath：地下室的浴室数量。
- FullBath 和 HalfBath：地上的浴室数量。
- BedroomAbvGr：地上卧室的数量。
- KitchenAbvGr：地上厨房的数量。
- KitchenQual：厨房的质量和条件。
- TotRmsAbvGrd：地上总房间数。
- Functional：房屋的功能性评级。
- Fireplaces：壁炉的数量。
- FireplaceQu：壁炉的质量。
- GarageType：车库类型。
- GarageYrBlt：车库建造的年份。
- GarageFinish：车库的装修程度。
- GarageCars：车库可容纳的汽车数。
- GarageArea：车库的总面积。
- GarageQual：车库的质量。
- GarageCond：车库的当前状态。
- PavedDrive：车道是否铺设。
- WoodDeckSF：木制平台的面积。
- OpenPorchSF 和 EnclosedPorch：开放和封闭门廊的面积。

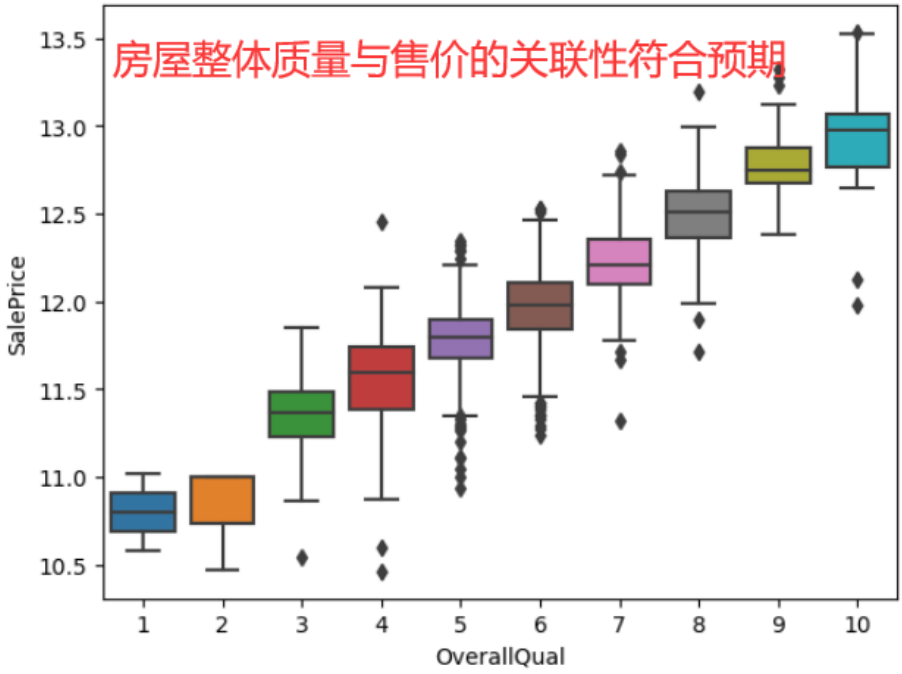
- 3SsnPorch：三季门廊的面积。
- ScreenPorch：屏幕门廊的面积。
- PoolArea：泳池的面积。
- PoolQC：泳池的质量。
- Fence：栅栏的类型和品质。
- MiscFeature：不在其他类别中的其他特征，如电梯、独立工作室等。
- MiscVal：其他特征的价值。
- MoSold：月份。
- YrSold：年份。
- SaleType：销售类型。
- SaleCondition：销售条件。

列名	含义
SalePrice	目标变量，表示房屋的销售价格。
MSSubClass	房屋的建筑类型，如一层、半层、双层等。
MSZoning	房屋所在地区的用途，如住宅区、商业区等。
LotArea	占地面积，单位是平方英尺。
Street	房屋所在街道的类型，是砾石路还是铺设路。
Alley	小巷通道的类型，有些房屋没有小巷通道。
LotConfig	房屋在街区内的位置，如内部、角落等。
YearBuilt	房屋建造的年份。
Neighborhood	房屋所在的社区，有些社区更受欢迎。

在训练过程中，我们通过数据分析的手段，首先分析各个特征与目标值（SalePrice）的相关性和相关系数，并处理填充异常值和空值，绘制了多张图表用以可视化显示结果，并最终选用数值型特征和类别特征构建了训练集用以输入到神经网络模型中。一部分数据分析的过程图片粘贴如下

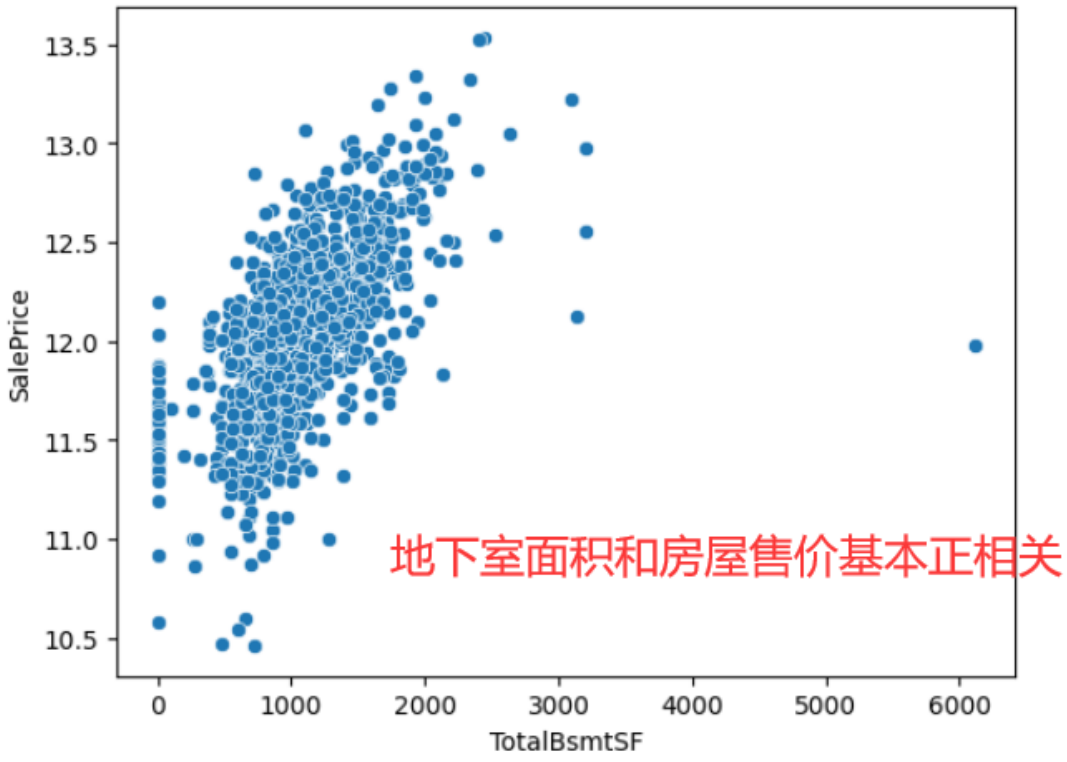
```
In [21]: sns.boxplot(data=train_data, x='OverallQual', y='SalePrice')
         #x轴表示房屋的整体质量  y轴表示房屋的销售价格
         #从图中看出，整体质量越高，销售价格也就越高
```

```
Out[21]: <AxesSubplot:xlabel='OverallQual', ylabel='SalePrice'>
```




```
4]: #绘制地下室总面积和房屋售价之间的关系，发现存在一定的正相关性
    sns.scatterplot(data=train_data,x='TotalBsmtSF',y='SalePrice')
```

```
4]: <AxesSubplot:xlabel='TotalBsmtSF', ylabel='SalePrice'>
```



打印相关性查看如下：

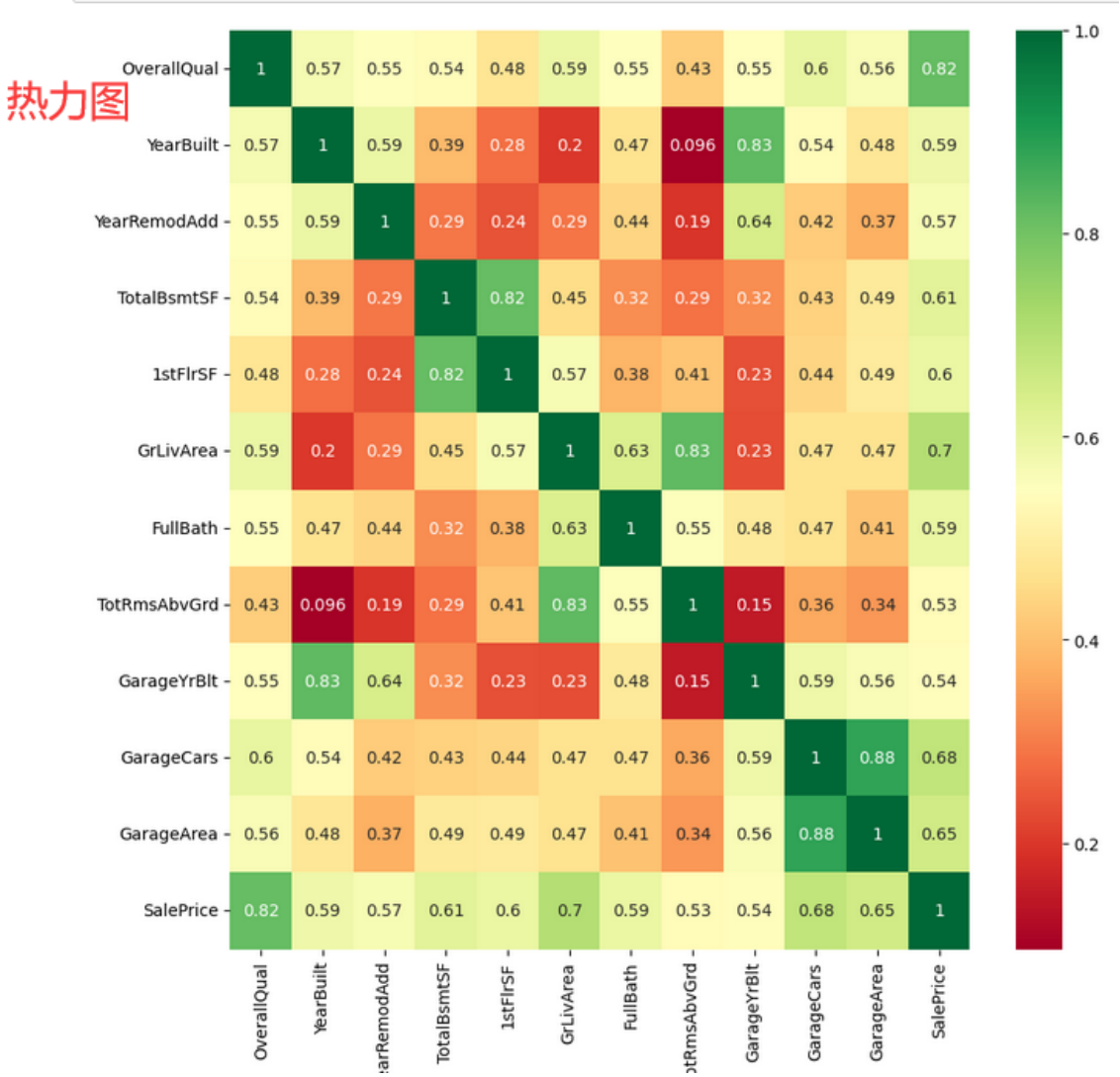
```
#查看训练集中每个特征与房价之间的相关性，并按照从高到低的顺序排序查看
#train_data.corr()计算了训练数据中的每个特征之间的相关性系数
#[1:]是去除房价自身的相关性系数 可以帮助更好的选择训练时所参考的特征
print(train_data.corr()["SalePrice"].sort_values(ascending=False)[1:])
```

```
OverallQual     0.817185
GrLivArea       0.700927
GarageCars      0.680625
GarageArea      0.650888
TotalBsmtSF     0.612134
1stFlrSF        0.596981
FullBath        0.594771
YearBuilt       0.586570
YearRemodAdd    0.565608
GarageYrBlt     0.541073
TotRmsAbvGrd    0.534422
Fireplaces      0.489450
MasVnrArea      0.430809
BsmtFinSF1      0.372023
LotFrontage     0.355879
WoodDeckSF      0.334135
OpenPorchSF     0.321053
2ndFlrSF        0.319300
HalfBath        0.313982
```

打印查看特征和目标的相关系数

各个特征的热力图如下

热力图



MNIST数据集

MNIST是一个手写数字识别数据集，由60000个训练样本和10000个测试样本组成。它被广泛用于深度学习中，并被认为是计算机视觉中的“Hello World”。

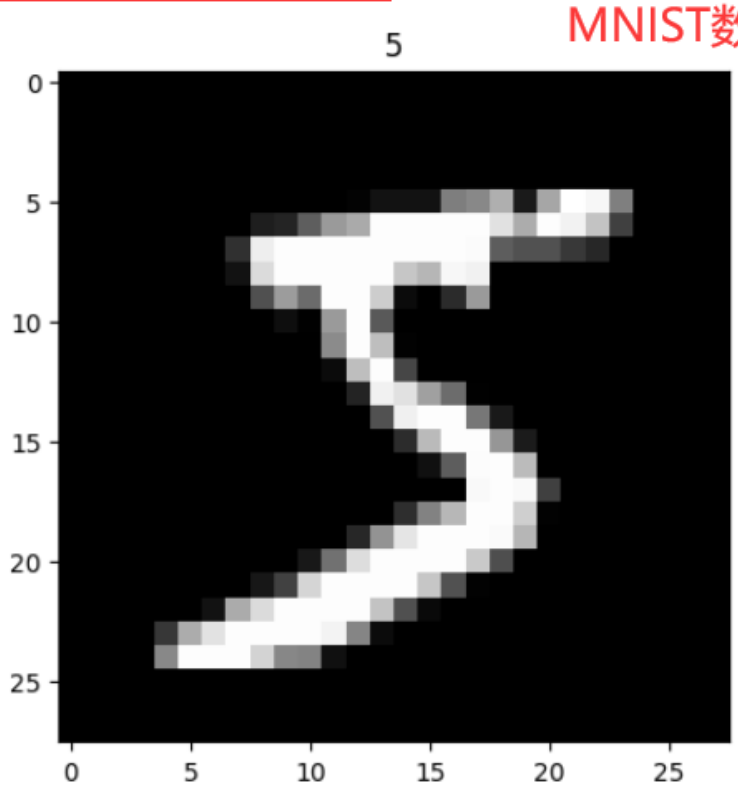
每张图片大小为28×28个像素，黑白单通道，每个像素的值都在0到255之间。其中，训练集包含10个类别的数字0-9的手写样本，每个类别各有6000个不同的样本。测试集同样包含10个类别的数字0-9的手写样本，每个类别各有1000个不同的样本。

```
# plot one example
print(train_data.data.size())      # (60000, 28, 28)
print(train_data.targets.size())   # (60000)
plt.imshow(train_data.data[0].numpy(), cmap='gray')
plt.title('%i' % train_data.targets[0])
plt.show()
```

由60000张28*28的图像组成

```
torch.Size([60000, 28, 28])
torch.Size([60000])
```

MNIST数据集



上交所股票数据集

000001SH_index.csv是上证综指（上海证券交易所综合股价指数）的股票数据集，包含了1998年至2020年的股票交易数据。

具体来说，该数据集包含了每天上证综指的开盘价、收盘价、最高价、最低价、成交量和成交额，以及对应的涨跌幅和换手率。其中，开盘价表示当天股票市场开市时的第一笔交易价格，收盘价表示当天股票市场收市时最后一笔交易价格，最高价和最低价则表示当天出现的最高和最低交易价格，成交量表示当天的交易量（以手为单位），成交额表示当天的交易金额（以万元为单位），涨跌幅表示当天的股价相对于前一天的涨跌幅，换手率则表示当天的股票换手率。

这个数据集可以用于许多金融分析和预测任务，例如股票价格预测、交易量分析、市场波动率估计等等。在使用数据集时，需要注意数据间可能存在复杂的关联结构，同时需要考虑到时间的因素，因为股票市场具有明显的时间序列性质，无论是对于特征提取还是建模都需要充分利用时间序列特点。

列名	含义
id	每条数据的唯一标识符
ts_code	股票代码
trade_date	交易日期
close	收盘价，在当天交易结束后的最后一笔成交价为收盘价
open	开盘价，在当天交易开始前的第一笔成交价为开盘价
high	最高价，在当天交易过程中的最高成交价为最高价
low	最低价，在当天交易过程中的最低成交价为最低价
pre_close	昨收价，即上一个交易日的收盘价
pct_chg	涨跌幅，即涨跌额与昨收价的比值
vol	成交量，即当天股票的成交量
amount	成交金额，即当天股票的成交金额

```
In [37]: data = pd.read_csv('./data/000001SH_index.csv')

In [38]: data.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5430 entries, 0 to 5429
Data columns (total 12 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   id          5430 non-null   int64
 1   ts_code     5430 non-null   object
 2   trade_date  5430 non-null   int64
 3   close       5430 non-null   float64
 4   open        5430 non-null   float64
 5   high        5430 non-null   float64
 6   low         5430 non-null   float64
 7   pre_close   5430 non-null   float64
 8   change      5430 non-null   float64
 9   pct_chg     5430 non-null   float64
 10  vol         5430 non-null   float64
 11  amount      5430 non-null   float64
dtypes: float64(9), int64(2), object(1)
memory usage: 509.2+ KB
```

上交所股票数据集

泰坦尼克存活数据集

Titanic - Machine Learning from Disaster竞赛的训练集是一个经典的机器学习数据集，包含了891名乘客的相关信息以及他们是否在船沉没后幸存下来的标签。

其中特征“Survived”为标签，表示乘客是否幸存。原数据集中共有891个样本，即891名乘客的信息，其中342名乘客幸存，549名乘客未幸存。

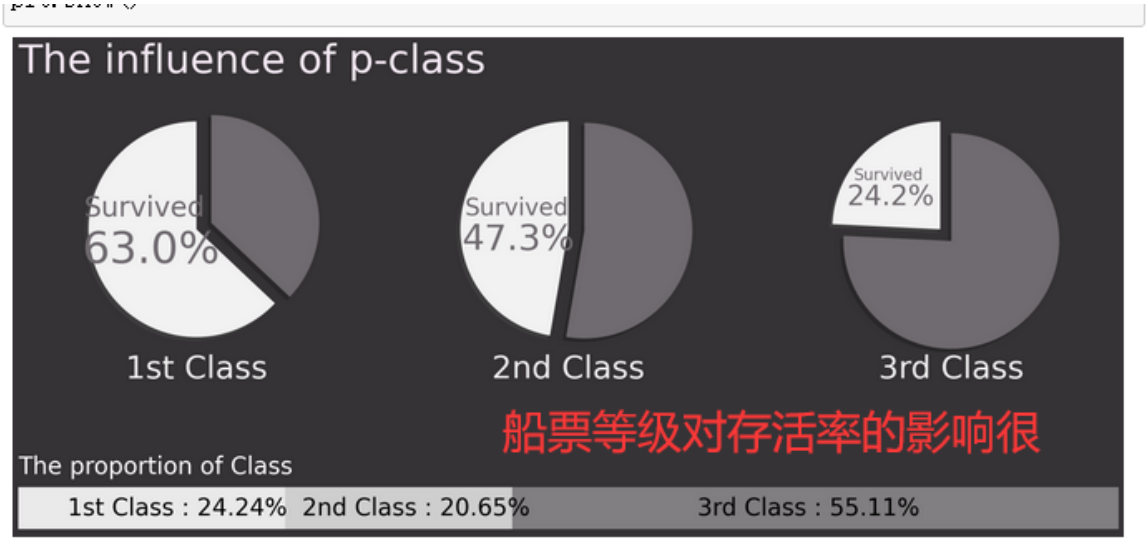
通过对该数据集进行分析和建模，可以通过已知乘客的相关特征来预测其是否幸存下来。这对于提高海难生还率、提高船舶安全等方面具有重要意义。

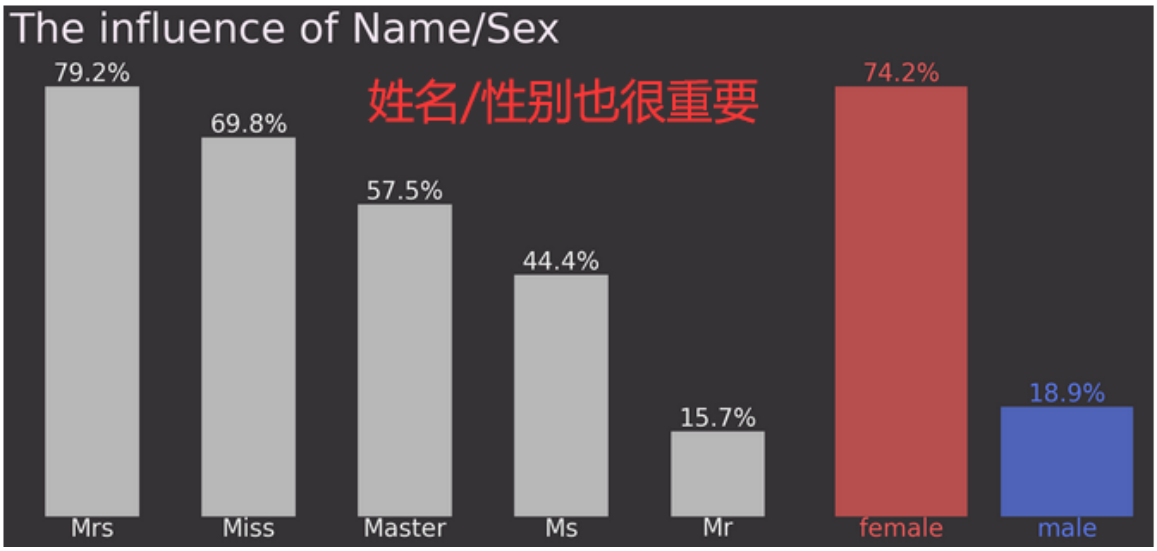
列名	含义
PassengerId	乘客ID
Survived	是否生还（0：遇难，1：生还）
Pclass	船票等级（1：一等舱，2：二等舱，3：三等舱）
Name	乘客姓名
Sex	乘客性别

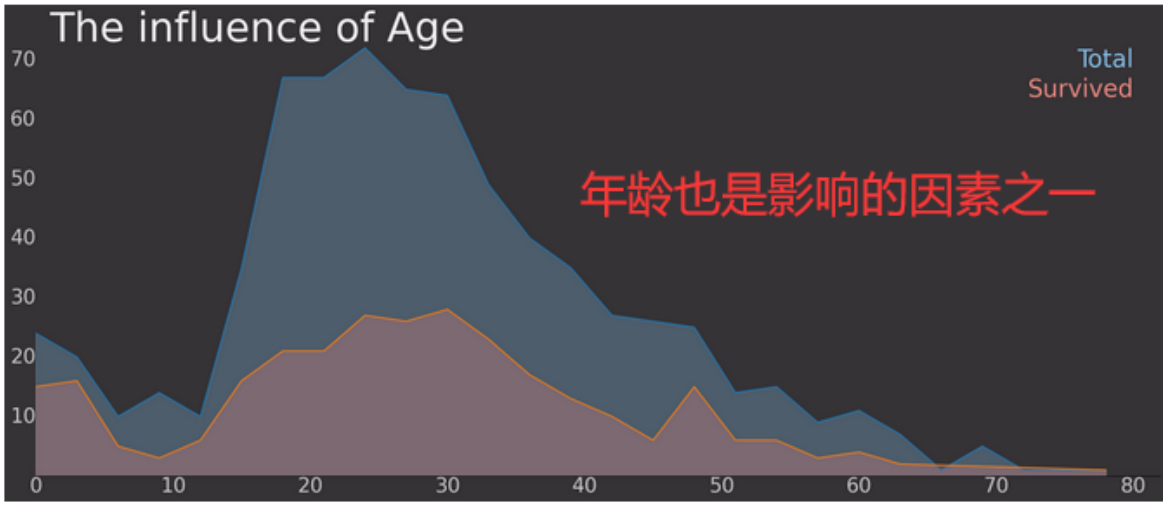
列名	含义
Age	乘客年龄
SibSp	配偶或兄弟姐妹的数量
Parch	父母或子女的数量
Ticket	船票号码
Fare	票价
Cabin	船舱号码
Embarked	登船港口（C: Cherbourg, Q: Queenstown, S: Southampton）

其中，Survived是要被预测的目标变量，其余11个特征均为作为模型训练的输入变量。

我们根据生活常识和SKlearn中的相关方法对数据集进行了分析和处理。








```
y = train['Survived']
X = train.drop('Survived',axis=1)
```

```
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler()
X = scaler.fit_transform(X)
test = scaler.transform(test)
```

```
X.shape #训练集
```

获取训练集和测试集

```
(891, 13)
```

训练集尺度为891*13

```
test.shape
```

那么网络的输入宽度即为13

```
(418, 13)
```

影视评论数据集

train.tsv是一个包含了14万条电影评论的文本文件，用于训练模型。该数据集包含了五列：PhraseId、SentenceId、Phrase、Sentiment和Aspect。

列名	含义
PhraseId	评论的唯一标识符
SentenceId	评论所在的句子的唯一标识符
Phrase	电影评论的文本内容
Sentiment	评论的情感极性标签，是一个5级分类标签，取值范围为0-4，分别表示“非常负面”、“稍微负面”、“中性”、“稍微正面”和“非常正面”

训练集包含了156060个短语片段以及短语片段所属的情感标签。竞赛数据集作为一个二分类任务，标签只有0或1，也就是说，标签为0的短语片段表示负面情感，标签为1的短语片段既可以表示中性情感也可以表示正面情感。因此，如果需要将该任务转化为三分类或更多分类任务，则需要进一步处理数据集。