Lab 4： 基于统计学习的泰塔尼克号生还预测

231880101 孙俊晖

1 引言

1.1 背景

1912年4月14日23时40分左右，当时世界上体积最庞大、内部设施最豪华的客运轮船—泰坦尼克号与一座冰山相撞后沉没，此次沉没事故为和平时期死伤人数最为惨重的一次海难，其残骸直至1985年才被再度发现，受到联合国教育、科学及文化组织的保护。Titanic数据集是一个关于此次沉没事故遇难者生存的数据集，包含乘客的性别、年龄、所在船舱、登船港口、票号、是否生还等信息。但由于部分乘客数据缺失，我们无从得知这部分乘客是否在那次灾难中幸存下来。

1.2 目标

使用所学的统计机器学习算法构建一个模型，使得该模型能够根据一名乘客的特征预测该乘客是否在那次灾难中幸存下来。

2 方法

2.1 技术栈

开发环境：Visual Studio Code 1.93.0

编程语言：Python 3.7.9 64-bit

2.2 读取数据

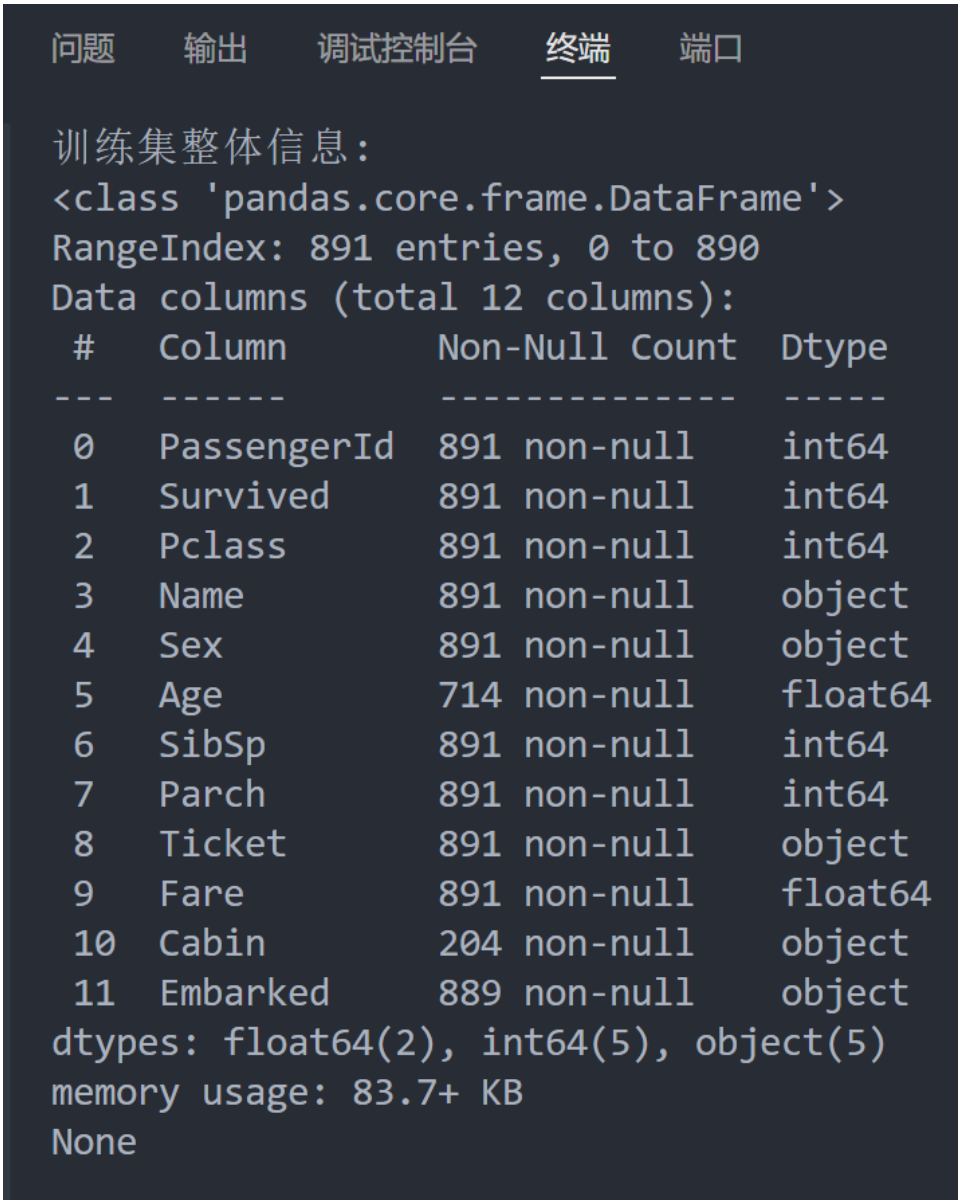
首先，利用pandas库读取Kaggle网站中提供的训练集**train.csv**和测试集**test.csv**

```
import pandas as pd
train = pd.read_csv('train.csv')
test = pd.read_csv('test.csv')
```

在读取完数据后，尝试查看训练集和测试集的信息，包括数据集的整体信息、描述性统计信息以及缺失值统计信息。

```
print("训练集整体信息:")
print(train.info())
print("\n测试集整体信息:")
print(test.info())
print("训练集描述性统计信息:")
print(train.describe())
print("\n测试集描述性统计信息:")
print(test.describe())
print("训练集缺失值统计:")
print(train.isnull().sum())
print("\n测试集缺失值统计:")
print(test.isnull().sum())
```

训练集整体信息：



测试集整体信息：

```
测试集整体信息:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 418 entries, 0 to 417
Data columns (total 11 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  418 non-null    int64
 1   Pclass       418 non-null    int64
 2   Name         418 non-null    object
 3   Sex          418 non-null    object
 4   Age          332 non-null    float64
 5   SibSp        418 non-null    int64
 6   Parch        418 non-null    int64
 7   Ticket       418 non-null    object
 8   Fare         417 non-null    float64
 9   Cabin        91 non-null     object
 10  Embarked     418 non-null    object
dtypes: float64(2), int64(4), object(5)
memory usage: 36.0+ KB
None
```

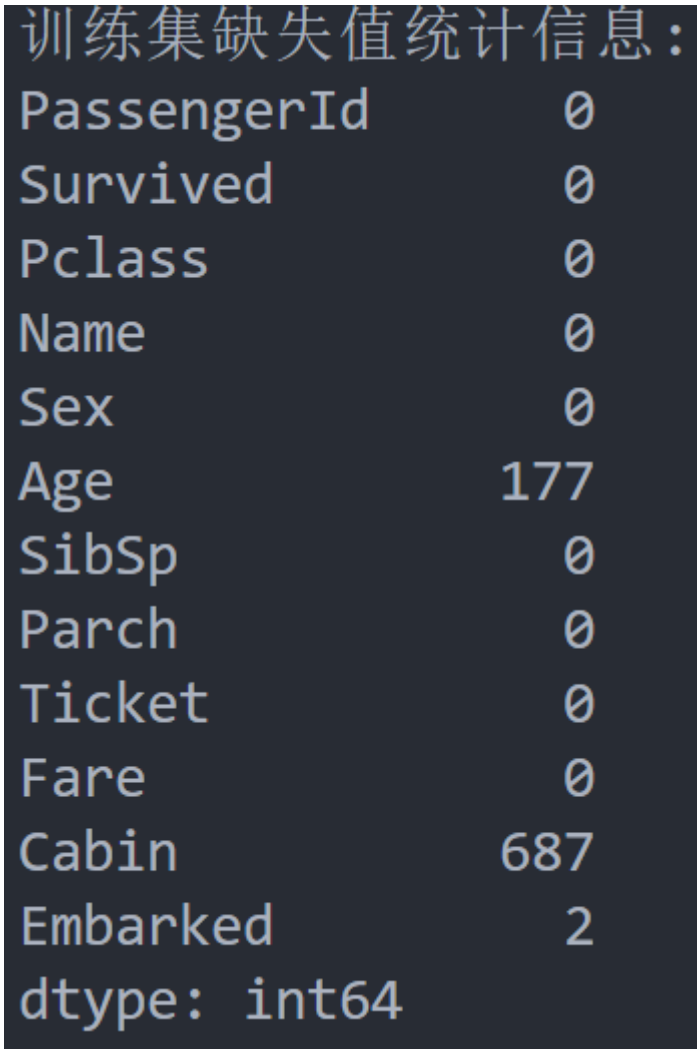
训练集描述性统计信息:

```
训练集描述性统计信息:
       PassengerId    Survived      Pclass         Age       SibSp       Parch        Fare
count   891.000000  891.000000  891.000000  714.000000  891.000000  891.000000  891.000000
mean    446.000000    0.383838    2.308642   29.699118    0.523008    0.381594   32.204208
std     257.353842    0.486592    0.836071   14.526497    1.102743    0.806057   49.693429
min       1.000000    0.000000    1.000000    0.420000    0.000000    0.000000    0.000000
25%     223.500000    0.000000    2.000000   20.125000    0.000000    0.000000    7.910400
50%     446.000000    0.000000    3.000000   28.000000    0.000000    0.000000   14.454200
75%     668.500000    1.000000    3.000000   38.000000    1.000000    0.000000   31.000000
max     891.000000    1.000000    3.000000   80.000000    8.000000    6.000000  512.329200
```

测试集描述性统计信息:

```
测试集描述性统计信息:
       PassengerId      Pclass         Age       SibSp       Parch        Fare
count   418.000000  418.000000  332.000000  418.000000  418.000000  417.000000
mean   1100.500000    2.265550   30.272590    0.447368    0.392344   35.627188
std     120.810458    0.841838   14.181209    0.896760    0.981429   55.907576
min     892.000000    1.000000    0.170000    0.000000    0.000000    0.000000
25%     996.250000    1.000000   21.000000    0.000000    0.000000    7.895800
50%    1100.500000    3.000000   27.000000    0.000000    0.000000   14.454200
75%    1204.750000    3.000000   39.000000    1.000000    0.000000   31.500000
max    1309.000000    3.000000   76.000000    8.000000    9.000000  512.329200
```

训练集缺失值统计信息:


训练集缺失值统计信息:
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

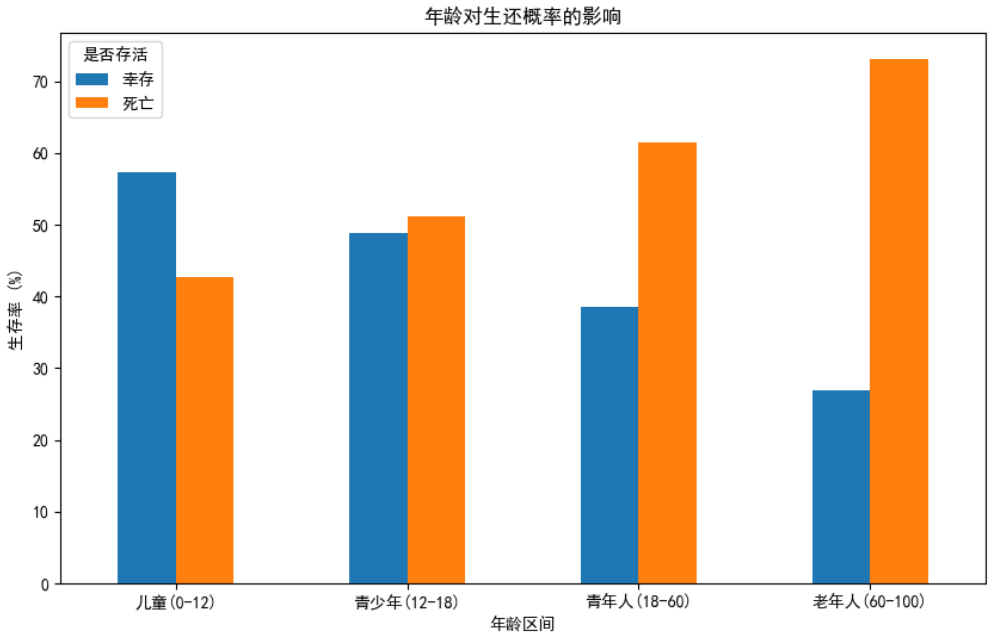
测试集缺失值统计信息：

```
测试集缺失值统计信息:
PassengerId      0
Pclass           0
Name             0
Sex              0
Age             86
SibSp            0
Parch            0
Ticket           0
Fare             1
Cabin          327
Embarked         0
dtype: int64
```

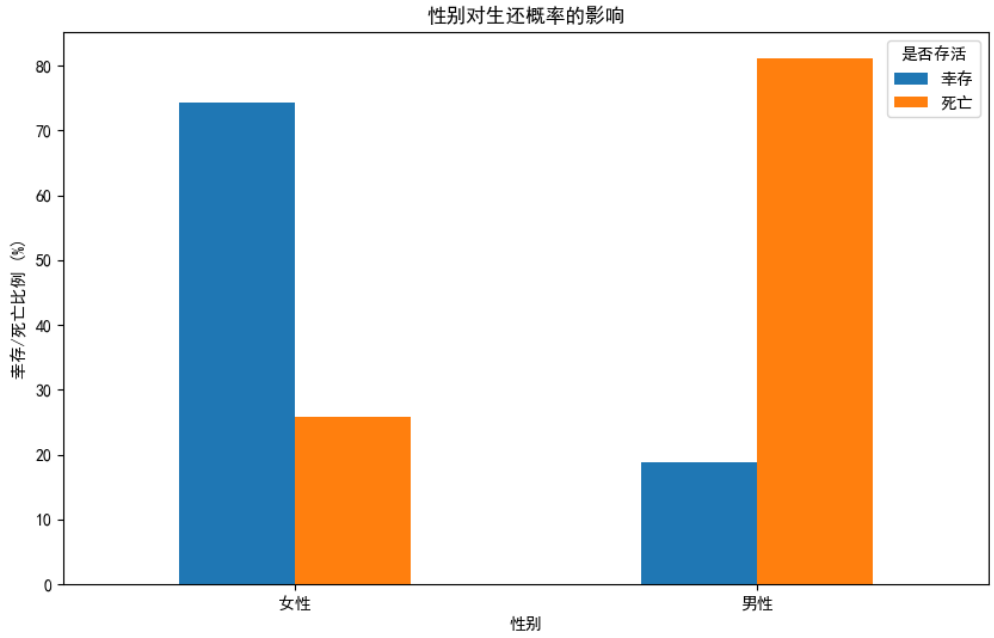
2.3 特征分析

接着，需要找出与生存相关的特征。这部分的工作已经在**Lab1**中做过了，在**Lab1**中我得出了一名乘客的**Age(年龄)、Sex(性别)、Pclass(舱位等级)**这三个因素对生存的影响很大，以下是我在Lab1中利用matplotlib可视化这三个因素对生还概率的影响。

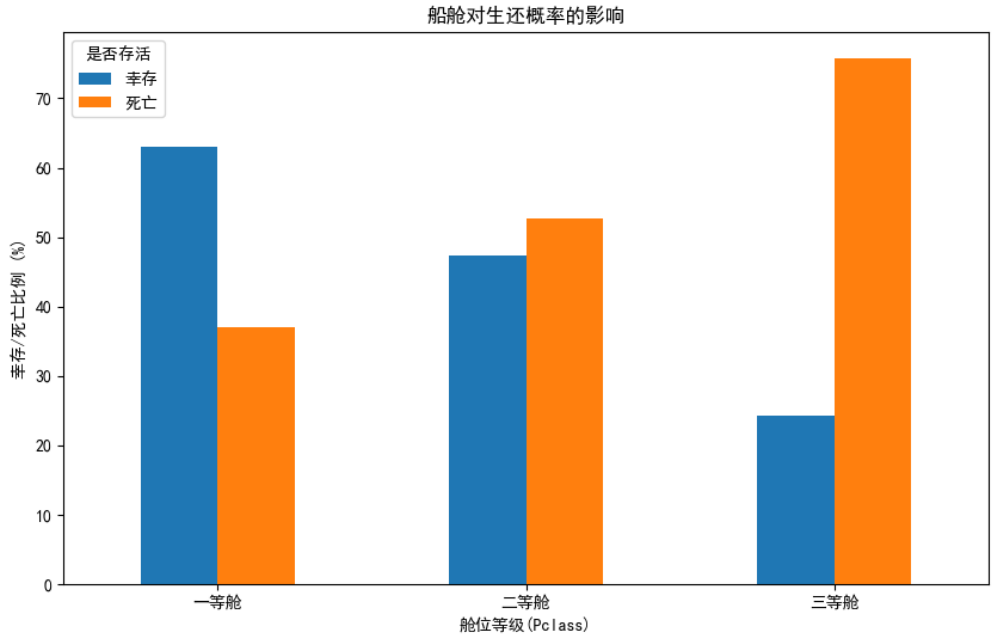
年龄对生还概率的影响：



性别对生还概率的影响：

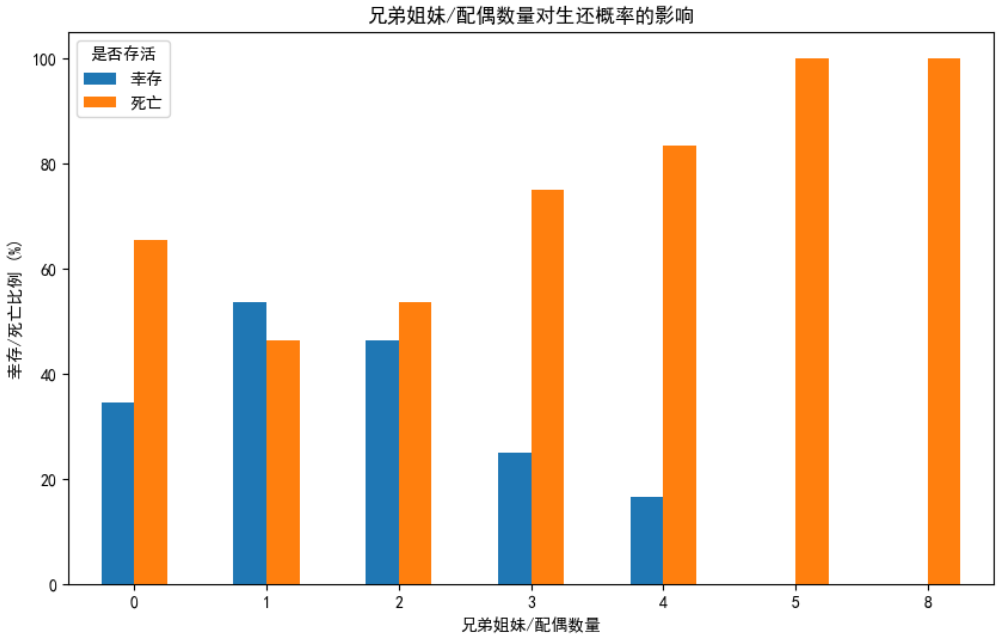


舱位等级对生还概率的影响：

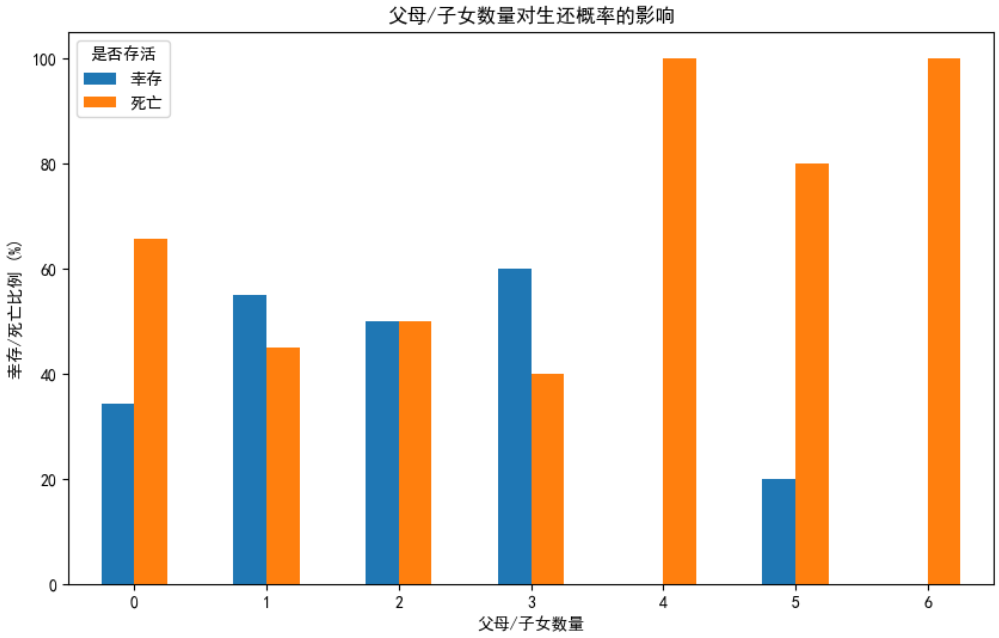


而**SibSp(兄弟姐妹/配偶数量)、Parch(父母/子女数量)、Fare(票价)、Embarked(登船港口)**这四个特征对生存亦有影响，但没有前三个特征的作用显著。以下是我利用matplotlib可视化这四个因素对生还概率的影响。

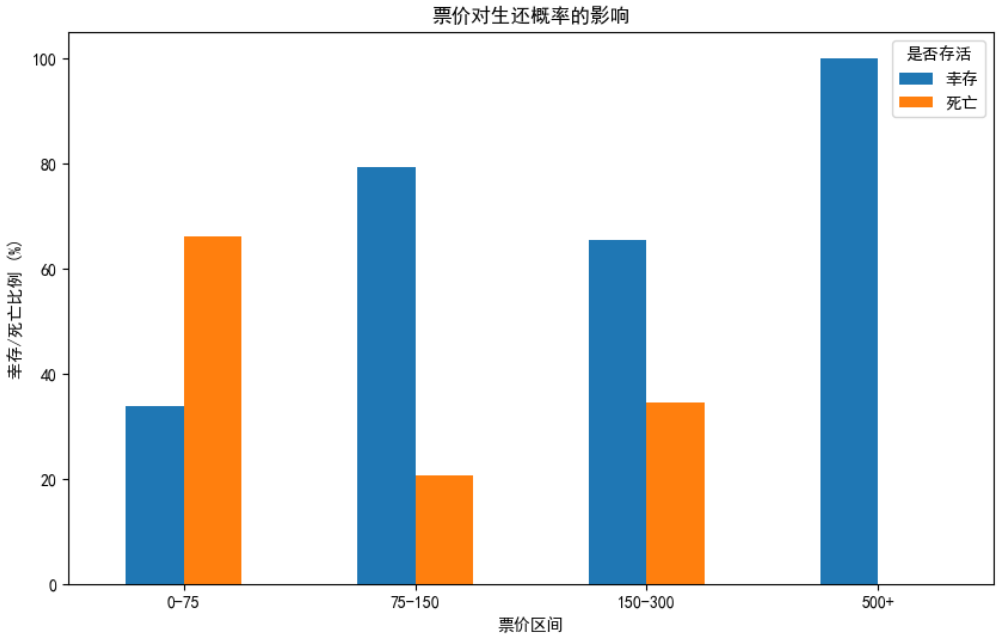
兄弟姐妹/配偶数量对生还概率的影响：



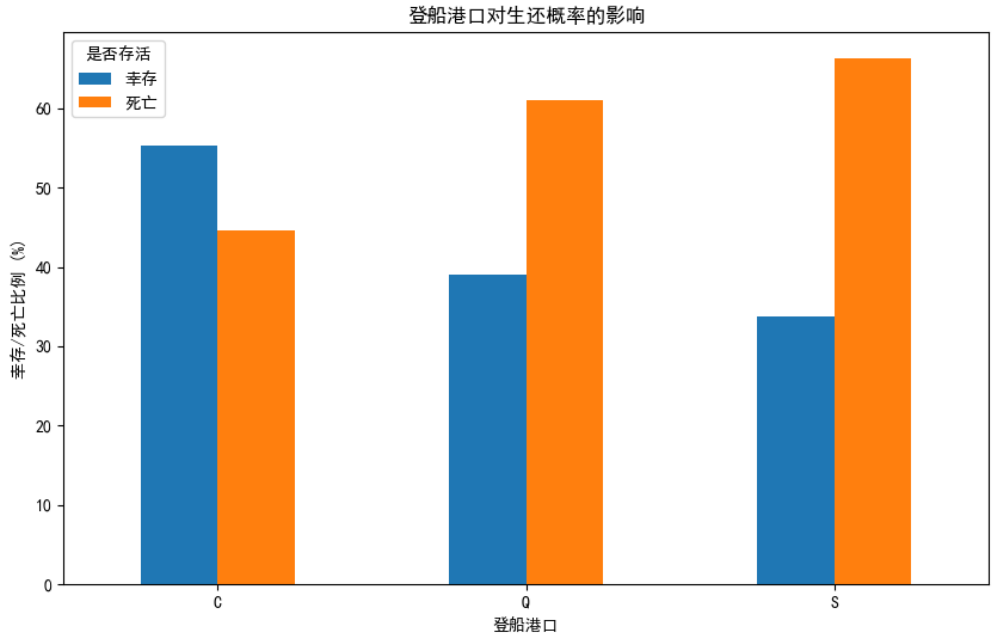
父母/子女数量对生还概率的影响：



票价对生还概率的影响：



登船港口对生还概率的影响：



2.4 缺失值处理

在查看完训练集和测试集的缺失值统计信息后，我发现在训练集和测试集中**Cabin**这一列都缺失了**接近80%**的数据，因此在特征分析中直接选择放弃分析Cabin对生还概率的影响，也不对这一列进行缺失值处理。而训练集和测试集中**Age**这一列缺失了**接近20%**的数据，不算太多，可以使用这一列已有数据的均值/中位数进行填充，最终我选择使用均值进行填充。训练集中Embarked这一列和测试集中Fare这一列中只缺失了**一两个**数据，我选择用已有数据的众数进行填充。

```
train['Age'].fillna(train['Age'].mean(), inplace=True)
test['Age'].fillna(test['Age'].mean(), inplace=True)
train['Embarked'].fillna(train['Embarked'].mode()[0], inplace=True)
test['Fare'].fillna(test['Fare'].mode()[0], inplace=True)
```

2.5 将分类数据转换为数值数据(编码)

在上述我分析的七个特征：**Age(年龄)、Sex(性别)、Pclass(舱位等级)、SibSp(兄弟姐妹/配偶数量)、Parch(父母/子女数量)、Fare(票价)、Embarked(登船港口)**中，Sex和Embarked的分类数据为非数值数据，由于大多数统计机器学习算法都是基于数值进行计算的，所以需要先将这两列的分类数据转换为数值数据，即进行"编码"。

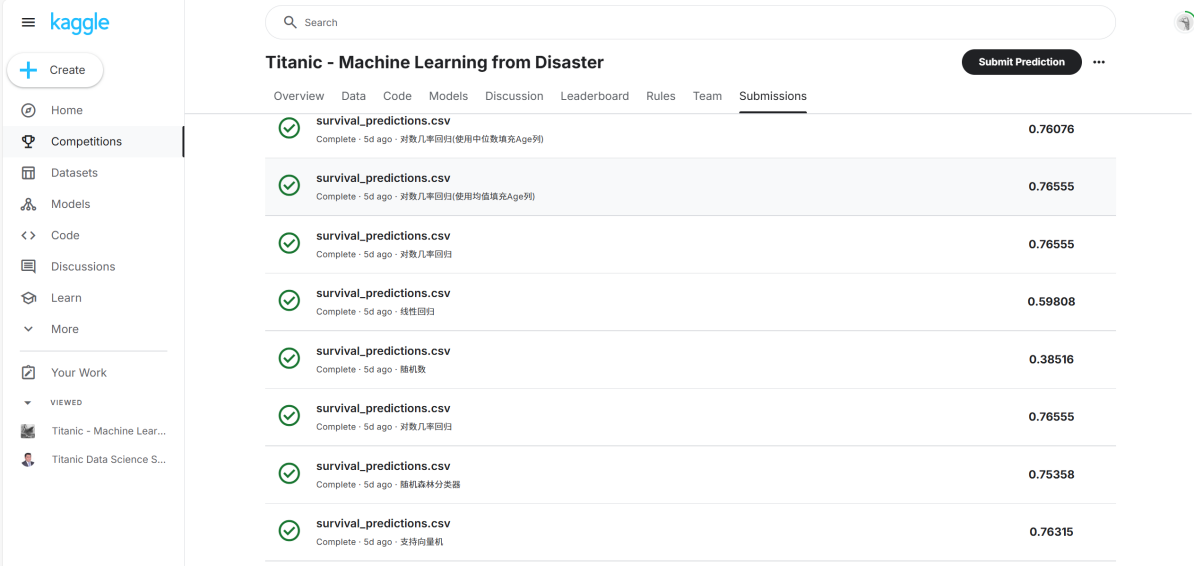
```
train['Sex'] = train['Sex'].map({'male': 0, 'female': 1})
test['Sex'] = test['Sex'].map({'male': 0, 'female': 1})
train['Embarked'] = train['Embarked'].map({'S': 0, 'C': 1, 'Q': 2})
test['Embarked'] = test['Embarked'].map({'S': 0, 'C': 1, 'Q': 2})
```

2.6 构建模型

尝试使用sklearn库中的模块，基于上述选择的七个特征构建模型。我分别尝试了使用Sklearn中**随机森林分类器、线性回归、支持向量机、对数几率回归**这几个模块训练模型，用训练好的模型预测测试集中乘客的存活情况，并将结果导出为csv文件在Kaggle平台上，最终使用对数几率回归模块训练的模型得分最高，为0.76555。

```
traindata = train[['Pclass', 'Sex', 'Age', 'SibSp', 'Parch', 'Fare',
'Embarked']]
trainlabel = train['Survived']
testdata = test[['PassengerId', 'Pclass', 'Sex', 'Age', 'SibSp', 'Parch',
'Fare', 'Embarked']]
#训练逻辑回归模型
clf = LogisticRegression(max_iter=1000)
clf.fit(traindata, trainlabel)
#预测存活情况
predictions = clf.predict(testdata.drop('PassengerId', axis=1))
#将结果导出为CSV文件
result = testdata[['PassengerId']].copy()
result['Survived'] = predictions
result.to_csv('survival_predictions.csv', index=False)
```

在Kaggle平台上的得分:



3 总结

在本次泰坦尼克号生还预测的实验中，我使用了统计机器学习的方法构建了一个预测模型。基于Lab1的工作，通过对数据集进行初步的探索性分析，我确定了影响乘客生存概率的七个关键因素：年龄、性别、舱位等级、兄弟姐妹/配偶数量、父母/子女数量、票价和登船港口。在处理缺失值时，我根据数据缺失的情况，采取了不同的策略。对于缺失率极高的特征（如Cabin），我选择了忽略；对于缺失率较低的特征（如Age和Embarked），我根据具体情况使用均值、中位数或众数进行了填充。在将分类数据转换为数值数据的过程中，我采用了映射的方法，将性别和登船港口等非数值特征转换为了数值特征，以便机器学习算法能够处理。在构建模型时，我尝试了多种机器学习算法，包括随机森林分类器、线性回归、支持向量机和对数几率回归，最终发现对数几率回归在处理该二分类问题时具有较好的性能。通过本次实验，我不仅加深了对统计机器学习算法的理解，还学会了如何在实际问题中应用这些算法。同时，我也认识到了数据处理和特征选择会严重影响模型性能。在未来的学习和实践中，我将继续探索更多的数据处理技术和机器学习算法，以提高模型的性能和准确性。

4 参考文献

1.https://blog.csdn.net/2302_81096429/article/details/142166561

2.https://blog.csdn.net/lsb2002/article/details/135756277