**Kaggle泰坦尼克号预测报告**

Author：顾志文

**1. 数据特征信息：**

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

**1.1 首先查看特征缺失状况，及特征的数据类型**

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

RangeIndex: 891 entries, 0 to 890

Data columns (total 12 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 PassengerId 891 non-null int64

1 Survived 891 non-null int64

2 Pclass 891 non-null int64

3 Name 891 non-null object

4 Sex 891 non-null object

5 Age 714 non-null float64

6 SibSp 891 non-null int64

7 Parch 891 non-null int64

8 Ticket 891 non-null object

9 Fare 891 non-null float64

10 Cabin 204 non-null object

11 Embarked 889 non-null object

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

memory usage: 83.7+ KB

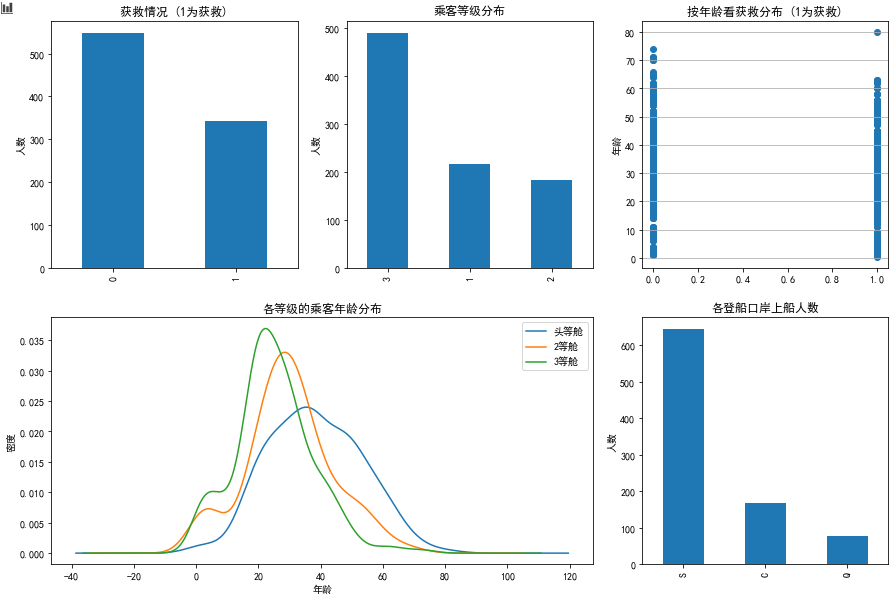
训练数据中总共有891名乘客，有些属性的数据不全，比如说：

Age（年龄）属性只有714名乘客有记录

Cabin（客舱）更是只有204名乘客是已知的

数值型变量mean字段表明，大概0.383838的人最后获救了，2/3等舱的人数比1等舱要多，平均乘客年龄大概是29.7岁，对数据有初步了解

**1.2 查看每个属性和最后的Survived之间有着什么样的关系**



在图上可以看出来:

被救的人300多点，不到半数；

\* 3等舱乘客灰常多；遇难和获救的人年龄似乎跨度都很广；

\* 3个不同的舱年龄总体趋势似乎也一致，2/3等舱乘客20岁多点的人最多，1等舱40岁左右的最多；

\*登船港口人数按照S、C、Q递减，而且S远多于另外俩港口。

一些特征工程方向：

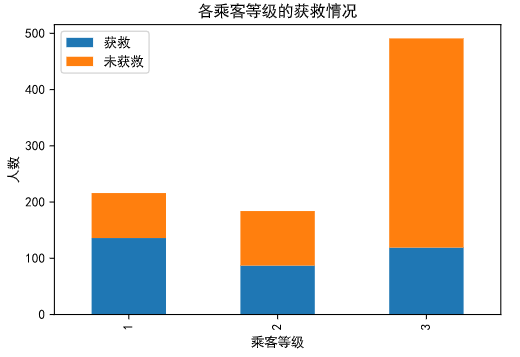
1 不同舱位/乘客等级可能和财富/地位有关系，最后获救概率可能会不一样

2 年龄对获救概率也一定是有影响的

3.和登船港口是不是有关系，也许登船港口不同，人的出身地位不同

**1.3查看这些属性值的统计分布。**

**1.3.1 看看各乘客等级的获救情况**

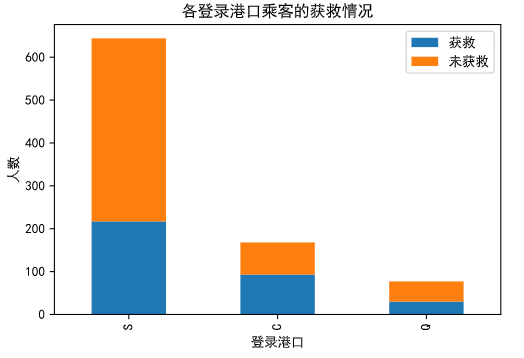


可以看出：

地位对舱位有影响，进而对获救的可能性也有影响

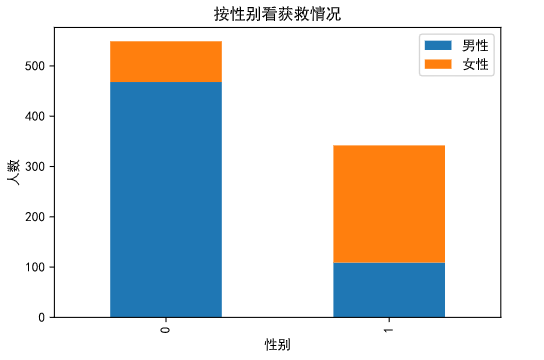
等级为1的乘客，获救的概率高很多。

**1.3.2 看看各登录港口的获救情况**



登录港口和获救之间没有明显特征

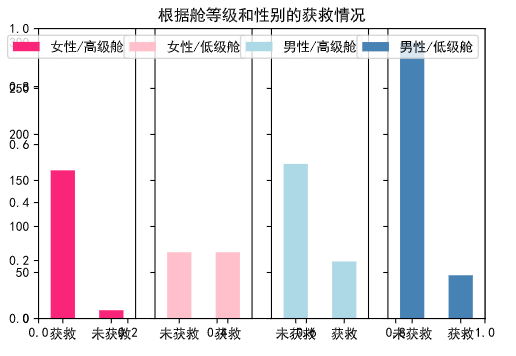
**1.3.3 看看各性别的获救情况**



可以看出：

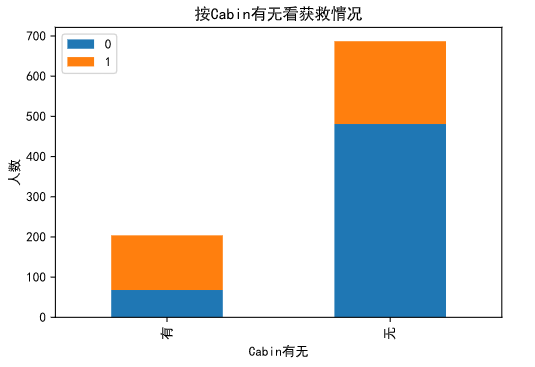
性别很重要，女性获救情况明显，无疑也要作为重要特征加入最后的模型之中。

**1.3.4 详细查看性别和其他特征的组合情况**



**1.3.5 ticket是船票编号，应该是unique的，和最后的结果没有太大的关系，不纳入考虑的特征范畴**

**1.3.6 cabin只有204个乘客有值，且分布离散，对其做有无特征处理**



有cabin记录的乘客survival比例稍高，把这个值分为两类，有cabin值无cabin值，

**2 特征工程**

**2.1 对于Age有少量缺失：**

用scikit-learn中的RandomForest来拟合一下缺失的年龄数据

**2.2 对类别特征做Onehot编码**

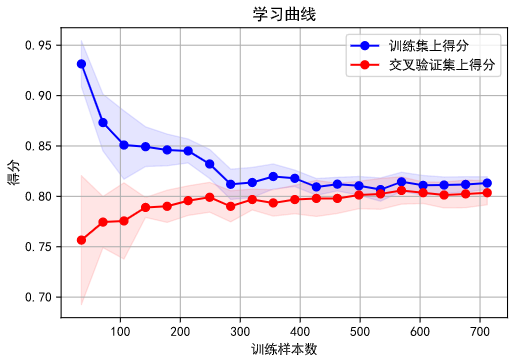
**2.3 对Age和Fare两个属性，进行归一化。这样可以加速logistic regression的收敛。**

**3初步建模，查看效果**

把需要的feature字段取出来，使用scikit-learn中的LogisticRegression建模。

单个弱分类器在测试集上准确率：0.76555

**4 判定一下当前模型所处状态(欠拟合or过拟合)**



在实际数据上看，我们得到的learning curve没有理论推导的那么光滑，但是可以大致看出，训练集和交叉验证集上的得分曲线走势还是符合预期的。

目前的曲线看来，我们的model并不处于overfitting的状态，因此可以再做些feature engineering的工作，添加一些新产出的特征或者组合特征到模型中。

**5挖掘新特征优化baseline系统，用交叉验证(cross validation)去验证**

在此之前，可以看看现在得到的模型的系数，因为系数表示这对target影响的重要性

columns coef

0 SibSp [-0.34969265843555536]

1 Parch [-0.1159483782914894]

2 Cabin\_No [-0.46681448665654995]

3 Cabin\_Yes [0.4667306801090533]

4 Embarked\_C [0.09104690422237205]

5 Embarked\_Q [0.08401916280240619]

6 Embarked\_S [-0.3498583736230197]

7 Sex\_female [1.3157828114750996]

8 Sex\_male [-1.3158666180225964]

9 Pclass\_1 [0.6176491672046102]

10 Pclass\_2 [0.2860850112582178]

11 Pclass\_3 [-0.903817985010329]

12 Age\_scaled [-0.5340503256369087]

13 Fare\_scaled [0.09220750646680881]

加入新特征，如：

* Name特征中，男性带某些字眼的(‘Capt’, ‘Don’, ‘Major’, ‘Sir’)可以统一到一个Title，女性也一样。
* Age属性不使用现在的拟合方式，而是根据名称中的『Mr』『Mrs』『Miss』等的平均值进行填充。
* 把堂兄弟/兄妹 和 Parch 还有自己 个数加在一起组一个Family\_size字段(考虑到大家族可能对最后的结果有影响)
* 单加一个Child字段，Age<=12的，设为1，其余为0（小孩获救率很高）
* Pclass和Sex组出一个组合属性

**6 Bagging 模型融合**

对于训练集，每次取训练集的一个subset，做训练，这样，虽然用的是同一个机器学习算法，但是得到的模型却是不一样的；同时，因为我们没有任何一份子数据集是全的，因此即使出现过拟合，也是在子训练集上出现过拟合，而不是全体数据上，这样做一个融合，可能对最后的结果有一定的帮助。