# 参与项目

## 海难沉船

一般流程是：1.数据可视化理解2.数据预处理和特征工程（缺失值，标准化(standardization/scaling)，规范化(normalization)）3.建模调参 4.评估整合

这个项目数据集有11维，涉及到年龄，兄弟姐妹，船舱号，票价，仓位等。

Cabin->fillna(‘X)

Embarked->map(纵向3维)->dummy(100)(横向3维)

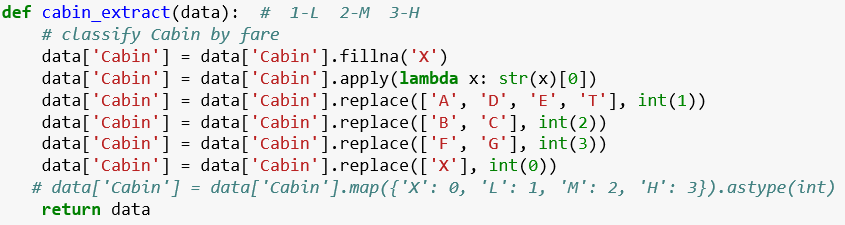
Age->title->mean->scaling

Fare->mean->scaling familysize, title

Sex->map（纵向2维）

1. 首先处理missing值， Age、Embarked、Cabin有缺失值。

对于Cabin（C85）两部分组成，提取字母，先用X填充缺失值，再把这A-X的字母分成0-3四类。



对于Embarked（S，Q，C）用众数填充，再换成0-2。

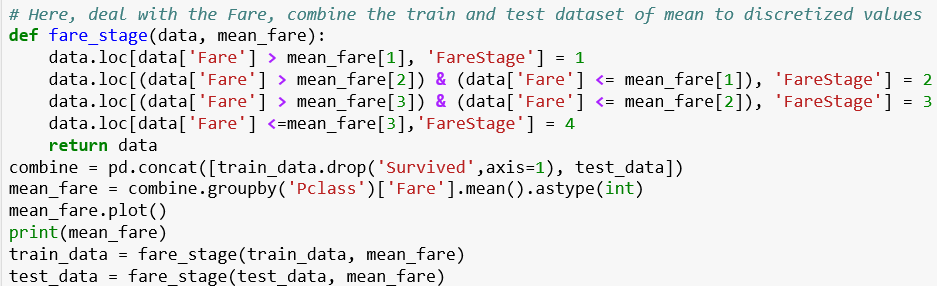


对于Age，后续添加了由name中提取的title属性后，用各自title的mean来填充。



1. 特征工程，创造新的属性 。

对于Fare票价属性，需要对它scaling，用mean来分箱。造FareStage属性1，2，3，4



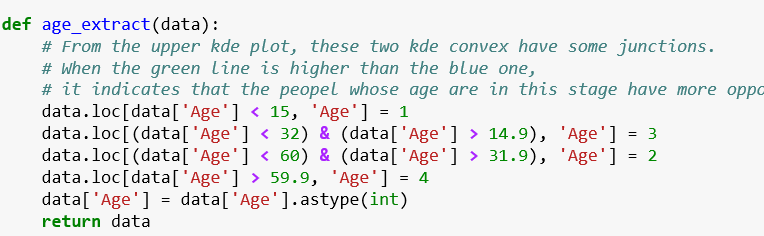
对于Name属性，提取title：Mrs，Mr，Miss等。有title（1-5）



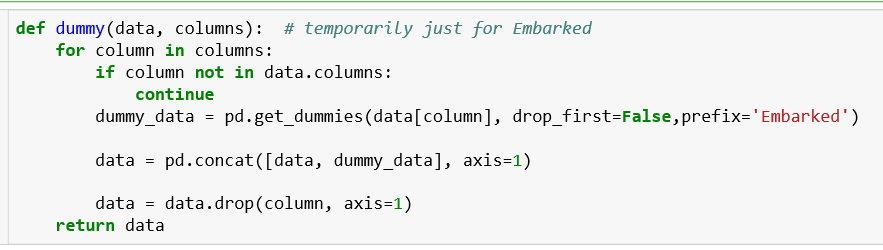
创造family\_info和Alone属性。



对于上面已经消除missing值的Age属性来说，需要scaling，分箱。



对于上面已经消除missing值的Embarked（0，1，2），需要用dummy二元化。



1. 完成特征工程后，建模调参。

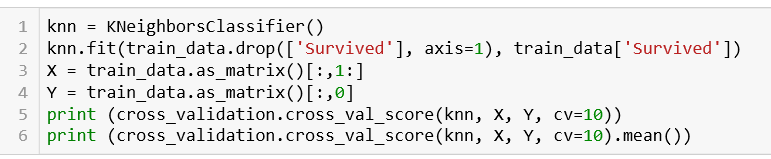
首先建立RandomForest模型，使用GridSearch法选参数。



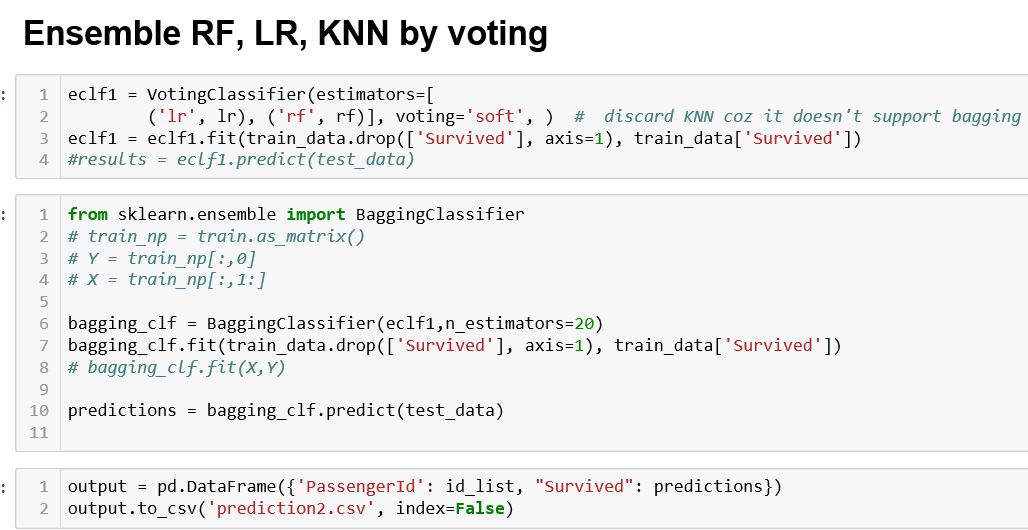
之后使用逻辑回归法并得到系数。



最后，建立KNN模型。



将三个模型使用ensemble方法，by voting



1. 画出learning curve。

