特征工程

泰坦尼克号生还者预测比赛（Manav Sehgal）：

**train数据**（891）中有12个特征属性如下： PassengerId、Pclass、Name、Sex、Age、SibSp、Parch、Ticket、Fare、Cabin、Embarked以及**标签**Survived。

其中Age缺失了177个值，Cabin缺失了687个值，Embarked缺失了2个值。

**test数据**（418）中有11个特征属性： PassengerId、Pclass、Name、Sex、Age、SibSp、Parch、Ticket、Fare、Cabin、Embarked。

其中Age缺失了86个值，Fare缺失了1个值，Cabin缺失了327个值。

**数值型的数据(int,float)**： PassengerId、Pclass、Age、SibSp、Parch、Fare、Survived。

**类别型的数据(object)**： Name、Sex、Ticket、Cabin、Embarked。

分别分析 **Pclass** 、 **Sex** 、 **SibSp** 、 **Parch** 与 **Survived** 之间的相关性，这些是从感觉上来的，毕竟仓位等级越高存活下来的可能性就越大，女性也应该会比男性更有优势存活下来。利用seaborn将数据可视化后来观察其他属性与Survived之间的关联性。

由于 **Ticket** 具有较高的重复率，而且可能跟Survived没有什么关联，所以这里就**删掉**。**（ Ticket => X ）**

**Cabin** 这个属性由于在训练和测试数据中都缺失了很多，所以**做删除处理**。**（ Cabin => X ）**

利用正则表达式将Name特征中的称呼抽出构成 **Title** 属性**作为一个新的特征**。由于称呼Master，Miss，Mr，Mis占有较多的数量，于是将这些抽出来的Title归为五类：Master，Miss，Mr，Mrs，Rare（将Mlle，Ms都归为Miss；将Mme归为Mrs；将其他的稀有少数称呼都归为Rare）。然后将提取出来的这五个属性利用map()方法映射成1，2，3，4，5。之后就可以保留Title字段，**删除 Name 属性**。同时**将 PassengerId也删除**。**（ Name PassengerId => Title ）**

对 **Age** 字段进行如下转化操作：

因为Sex和Pclass和Age之间的关系比较近，所以利用Sex和Pclass的组合字段所对应的Age的**中位数**来对缺失的数据进行填充。例如：Sex=1 & Pclass=1 =》Age.medain() 去填充缺失数据中Sex=1 & Pclass=1对应的Age值。这里Sex和Pclass有六种组合{（0，1），（0，2），（0，3），（1，1），（1，2），（1，3）}。这是Age已经没有空值了，之后再对Age的年龄段进行cut()为五段，每段依次赋值为0，1，2，3，4。

创建一个新的属性为 **FamilySize** ，其等于SibSp+Parch+1。再根据FamilySize创建另一个新属性 **IsAlone** 是否为单身。如果FamilySize=1，则IsAlone=1表示单身。这时就可以**将FamilySize、SibSp、Parch删掉，保留IsAlone**。**（ FamilySize SibSp Parch => IsAlone ）**

再创建一个 **Age\*Pclass**  特征。

对 **Embarked** 缺失的属性利用其**众数**进行填充。mode()方法可以取众数。再将Embarked下的种类S、C、Q进行数值型转化为0、1、2。map()方法实现。

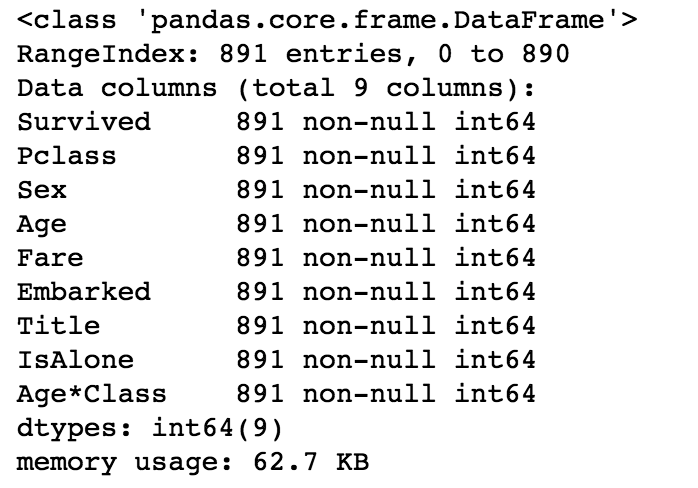
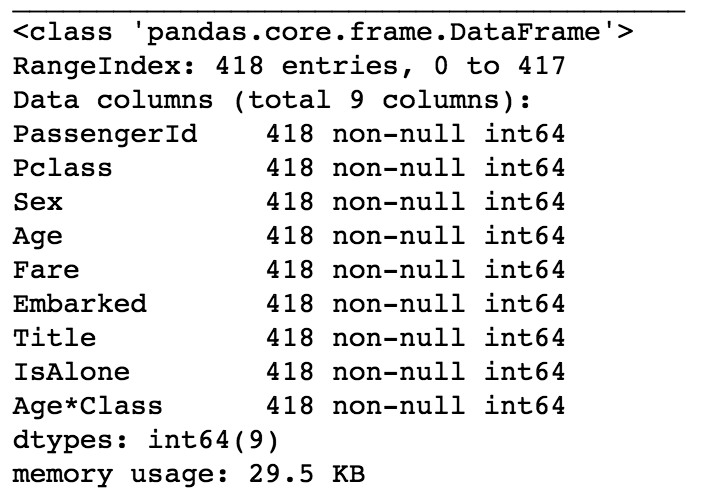
对 **Fare** 属性中缺失的那一个值用**中位数**填充。然后对Fare属性进行qcut()划分为4段，每个区间段将其数字化为0、1、2、3。

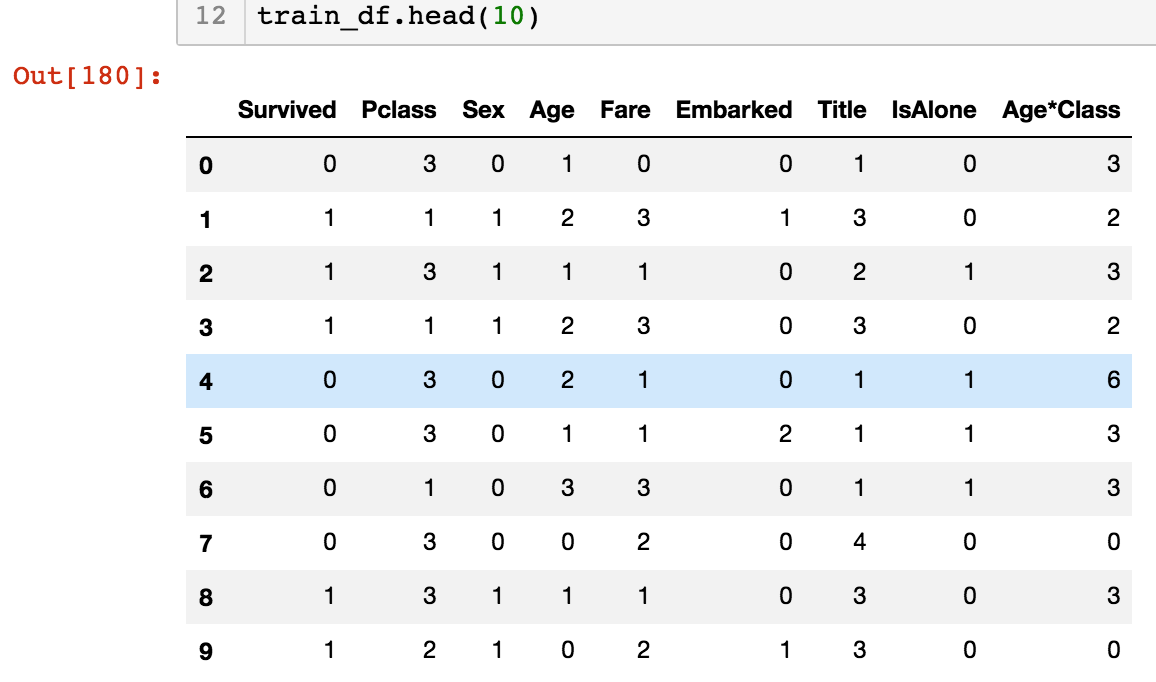
将 **Sex** 字段进行类别型转数值型操作，即用map()函数将female->1,male->0。

**到此，数据清晰和特征工程阶段告一段落。**

**接下来就是模型的训练和优化以及评估。**

最后得到所有数据都是int64的训练和测试数据。其中训练集包含8个特征：Pclass、Sex、Age、Fare、Embarked、Title、IsAlone、Age\*Class。测试集包含同样的特征。

train：test：



泰坦尼克号生还者预测比赛（LD Freeman）：

1、数据科学的框架

（1）定义问题

（2）收集数据

（3）准备数据

（4）探索性分析

（5）训练模型

（6）验证模型

（7）优化

用中位数填充Age的缺失值（train中缺177，test中缺86）

用众数填充Embarked的缺失值（train中缺2）

用中位数填充Fare的缺失值（test中缺1）

删除train中的PassengerId、Cabin、Ticket。

创建FamilySize = SibSp + Parch + 1

创建IsAlone 根据FamilySize来

从Name中用正则匹配出Title特征

将Fare划分成四段从而产生FareBin特征

将Age划分成五段从而产生AgeBin特征

到此train中有：Survived、Pclass、Name、Sex、Age、SibSp、Parch、Fare、Embarked、FamilySize、IsAlone、Title、FareBin、AgeBin共14个特征。

test中有：PassengerId、Survived、Pclass、Name、Sex、Age、SibSp、Parch、Fare、Cabin、Embarked、FamilySize、IsAlone、Title、FareBin、AgeBin共16个特征。

将Sex、Embarked、Title、AgeBin、FareBin通过LabelEncoder()进行特征编码fit\_transform()之后得到Sex\_Code、Embarked\_Coder、Title\_Coder、AgeBin\_Coder、FareBin\_Coder 这5个新特征。

train中添加5个特征之后就有19个特征了，

test中添加5个特征之后就有21个特征了。

**train1\_x：**

Sex\_Code Pclass Embarked\_Code Title\_Code SibSp Parch Age Fare

105 1 3 2 3 0 0 28.0 7.8958,

**train1\_x\_bin:**

Sex\_Code Pclass Embarked\_Code Title\_Code FamilySize AgeBin\_Code FareBin\_Code

1 3 2 3 1 1 0

**train1\_x\_dummy:**

Pclass SibSp Parch Age Fare FamilySize IsAlone Sex\_female

3 0 0 28.0 7.8958 1 1 0

Sex\_male Embarked\_C Embarked\_Q Embarked\_S Title\_Master Title\_Misc 1 0 0 1 0 0

Title\_Miss Title\_Mr Title\_Mrs

0 1 0

alg.fit(data1[data1\_x\_bin], data1[Target])

使用的是这个数据进行的训练。