**Titanic预测竞赛思路与历程**

1. **问题分析**

**1、问题是什么？**

泰坦尼克号存活预测是典型的有监督二分类问题，标签为用0、1表示的。

**2、可用模型有哪些？**

单模型：Logistics Regression、Decision Tree、SVM、Naïve Bayes

集合模型：Bagging、Randomforest、ExtraTrees（和随机森林一样，随机选取部分特征做决策树，但不同于随机森林选取每个特征的最佳分界值，ExtraTrees随机选取一些分界值，并从这些分界值中选取最优。）、Adaboost、GradientBoosting（GBDT）、Xgboost

**3、数据是否均衡？**

死亡549，存活342，在一个数量级上，数据基本可以当做均衡数据处理（为了验证数据的非均衡性对模型准确率影响到底有多大，后续在进行交叉验证时，分别选用f1 score和accuracy两个指标进行验证，结果相差甚微。）

1. **数据清洗**

**1、缺省值如何处理？**

1）对于缺省值太多的特征，可以考虑直接去掉该特征

2）可以考虑将包含缺省值的特征转化为类别类特征：缺省、非缺省（如Cabin特征）。

3）填充缺省值：

使用平均数、中位数、众数填充；

使用非缺省数据训练预测模型，预测缺省值（如使用了Randomforest模型预测缺省Age）

多重插补法预测缺省值（R语言mice）

**2、如何挖掘有效特征？**

在Titanic例子中，Sex是一个非常重要的特征（仅用sex做预测，准确率可达），模型在做预测时，倾向于将所有男性归类为非存活，而将女性归类为存活，如果一个家庭中有男性存活，则很大概率该家庭中所有成员都存活，反之亦然。据此可挖掘出两个新的特征：'perishing\_female\_surnames'和'surviving\_male\_surnames'，相当于对Sex这个重要特征做了一个校正。

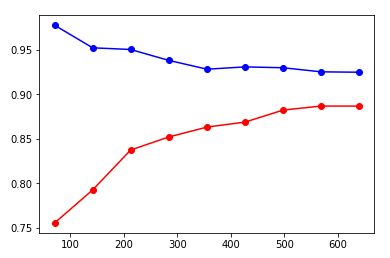
**3、数据如何转换？**

Pandas里面的categorical函数和sklearn里面的OneHotEncoder函数都可以将类别类特征数字化，但对于连续型特征（Age和Fare），是否应该进行一个分箱操作将其转换为类别特征？如果人为做分箱，并不能保证找到最优分箱数和分界阈值，而以决策树为基模型的模型会遍历并找到最优分界值，所以人为分箱一定程度上掩盖了特征的信息。而如果不做分箱，在训练数据过少的情况下，模型并不能完全学习到连续特征的所有信息，会导致模型过拟合；另一个问题是决策树的创建过程中，有这样一个特性，即特征分类数越多，则用该特征算出的信息增益越大（虽然信息增益率能避免此问题，但sklearn包中只有‘entropy’和‘gini’两个选项），该特征也容易以压倒性的优势成为最重要特征（xgboost模型很显著地表现出这种特点。）

1. **模型性能**

使用Randomforest模型进行预测时，尝试各种特征组合以及调参都不能提高模型性能，准确率始终在0.77左右。直到画出模型的learning curve，才找到模型性能无法提高的症结。

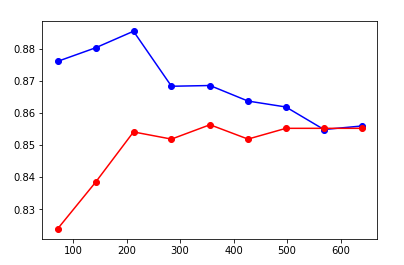
起初，learning curve如下：



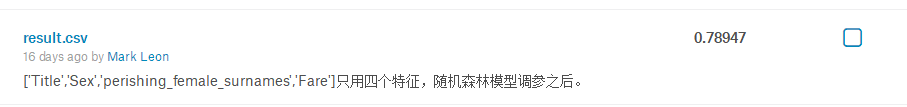
可以明显看到模型的train score和test score没有交汇，说明在有限的数据量下模型并没有得到充分训练，仍然处于过拟合状态。训练模型好比学生学习考试，最终考试分数低无非两个原因：一是学习内容太难，二是学习时间太短，导致最终没有完全掌握所学。如此，解决办法也一目了然：或者降低学习内容的难度，或者增加学习的时间：

**1、降低学习难度（简化特征数据）**

简化数据特征有两种途径：一是做特征选择，二是将连续性数据特征做简单的离散化（回答了之前连续特征是否要分箱的问题）。将“Fare”和“Age”简单分箱，只选取前4个重要特征训练模型并调参之后，learning curve如下，train score和test score完美拟合。

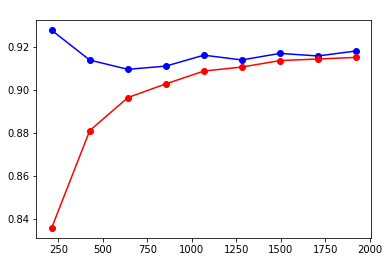


提交预测结果，准确率提升至0.7894。

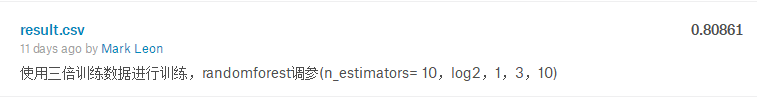


**2、增加学习时间（更多训练数据）**

虽然简化数据特征提高了准确率，但是这种做法也丢掉了太多重要的特征，致使模型过于简单。如果想要更多特征参与的情况下仍然让模型不过拟合，就只能增加更多的训练数据。将训练数据复制三份，前后衔接作为新的训练数据，选取前8个重要特征，训练模型并调参后，train score和test score很好地拟合：

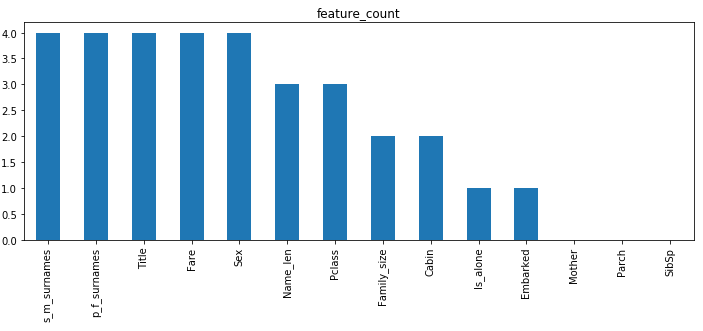


提交结果，准确率进一步提升至0.8086。

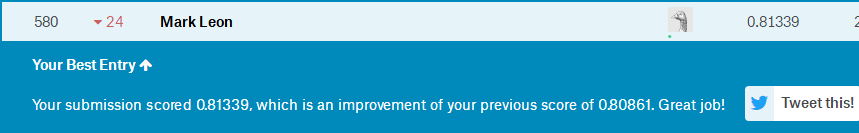


1. **投票机制**

多模型投票可以进一步提升模型性能，这里选取了RandomForestClassifier、XGBClassifier、GradientBoostingClassifier和ExtraTreesClassifier四个集合模型进行投票。每个模型都有不同的特征重要度排序，最后投票时如何做特征选择呢？这里采用了特征投票机制，即如果该特征在一个模型中重要度排前8位，则该特征计数加一，最终的计数表示该特征在几个模型中重要度能排到前8，结果如下：



最后依据该计数排序，选择前9位的特征进行训练，在保证每个模型的train score和test score都能很好拟合的情况下，建立投票模型VotingClassifier，预测并提交结果，准确率上升至0.81339，排名580/9593。



1. **总结**

Titanic预测竞赛的关键症结就在数据量太少，提高模型性能表现的关键就在于简化模型或者增加数据量，Kaggle网站的Kernel里面有些脚本在不做特征选择，不增加训练数据量的情况下，仍能得到较高的准确率，是因为其通过调节一些模型参数而使得某些关键特征的重要度异常高，弱化了其它特征的贡献，本质上也是进行了特征选择，只不过是通过调节模型参数实现了这一点。