机器学习大多数时候是在进行特征工程，特征决定 了机器学习预测效果的上限，而算法只是不断地去逼近这个上限而已

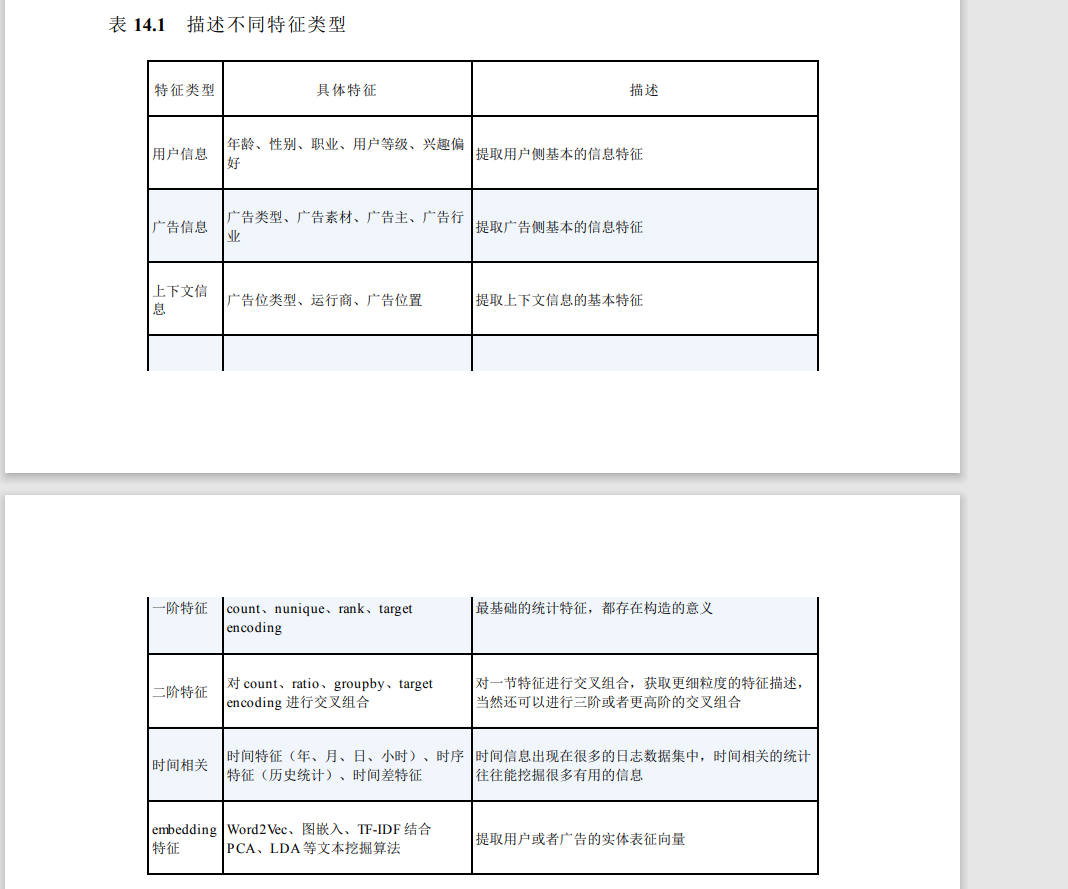
点击率 = 点击人数（次）/ 曝光人数（次）

迁移学习的方式

广告是否被点击的主导因素是用

户，其次才是广告信息。所以我们要做的是充分挖掘用户以及用户行为的信息，然

后才是广告主、广告等信息。本赛题的评价指标为对数损失。标签分布不 平衡的问题进行优化处理，通常是对负样本进行欠采样，对正样本进行过采样



数据分析：

数据清洗:清洗脏数据，string的数据利用自然数编码（有排序关系）或者one-hot(无排序关系) 使用独热编码可以让不同的分类处在“平等的地位”，不会因为数值的大小而对分类造成影响。

count、nunique、ratio 这三类是竞赛中类别特征经常使用的构造方式。count（计数特征）用于统计类别 特征的出现频次。nunique 和 ratio 的构造相对复杂一些，经常会涉及多个类别特 征的联合构造，例如在广告点击率预测问题中，对于用户 ID 和广告 ID，使用 nunique 可以反映用户对广告的兴趣宽度，也就是统计用户 ID 看过几种广告 ID； 使用 ratio 可以反映用户对某类广告的偏好程度，也就是统计用户 ID 点击某类广 告 ID 的频次占用户点击所有广告 ID 频次的比例。当然，这也适用于其他问题，比 如恶意攻击、反欺诈和信用评分这类需要构造行为信息或分布信息描述的问题。

比较强的特征是user，构造所有user与id的组合

本文为kaggle上一位选手分享的xgboost调参经验的翻译。方便对xgboost模型参数的快速调试。  
原文：https://www.kaggle.com/c/bnp-paribas-cardif-claims-management/forums/t/19083/best-practices-for-parameter-tuning-on-models  
数据的划分：一般从训练集里划分20%作为验证集，简易的调参不做交叉验证，因为交叉验证实在是耗费时间。  
模型参数的初始化：

模型类型：根据问题（二分类，多分类，回归等）选择合适的模型类型  
初始参数：eta= 0.1, depth= 10, subsample=1.0, min\_child\_weight = 5, col\_sample\_bytree = 0.2(depends on feature size)  
nround：xgboost的轮数，直接设置为10000，若error开始不断变高（或auc不断变低），直接中断程序运行即可  
调参  
1)先调整depth，一般depth与其他参数关联性不大，故第一个调整。首先将其初始化为10，训练模型观察最优结果。再将其调整为8，若效果变差则调整为12。用这种方法逐步接近较好的参数。  
2）然后调整subsample，初始化为1，观察结果。调整为0.8，若变好则继续下调一点看看，否则则改回0.9。若0.9的效果还不好则改回1  
3）用与depth相同的方法调整min\_child\_weight  
4）然后调整col\_sample\_bytree  
5）最后将eta设置为0.05（或者更低一点），训练模型并观察，记录最优的训练轮数（num\_round）——即错误率开始从不断下降转为不断上升的时候

具体参数对应含义可查：<https://blog.csdn.net/fuqiuai/article/details/79495910>

XGBoost 还可以有效地处理缺失值，训练时对缺失值自动学习切分方向。基 本思路是在每次的切分中，让缺失值分别被切分到决策树的左节点和右节 点，然后通过计算增益得分选择增益大的切分方向进行分裂，最后针对每个 特征的缺失值，都会学习到一个最优的默认切分方向。