EDA统计

- train_transaction
  - 590540条数据，除去transactionID（每条数据的识别ID,是唯一存在的），以及isFraud这个y以外有392个特征
  - isFraud不均衡，56W多未欺骗，2W多欺骗
  - 372列特征存在缺失值
  - 339列特征都是V开头的，匿名特征
  - 6列card开头的特征
    - card6训练集缺失2‰，测试集缺失6‰
取值为charge card, credit, debit, debit or credit, 其中charge card 和 debit or credit 分别有15，30条数据，都是未欺骗
debit欺骗占比较高，40倍左右，credit欺骗占比较低，在13倍左右
    - card4 训练集缺失3‰，测试集缺失6‰。
american express欺骗占比在38倍，discover欺骗在12倍，mastercard和visa30倍左右
    - card5： 训练集缺失7‰，测试集缺失9‰
取值在100-240，其中220以上取值异常多
    - card3 训练集缺失3‰，测试集缺失6‰
取值100-220左右，其中150有52W多，185有接近6W，其余很少
    - card2 训练集差点2%，测试集差点2%，取值从100-600左右，分布不明显，多峰多谷
    - card1 未缺失，取值在2000-18000，分布不明显，多峰多谷（不过相对card2要平稳）
  - 14列C开头的特征
  - 15列D开头的特征
  - 9列M开头的特征
  - addr1,addr2
    - 两者在训练集测试集都是同时缺失，训练集缺失11%，测试集上缺失13%
addr1在训练集上取值从100-500多，分布没有什么特征，
addr2在训练集上取值主要在60范围左右，但87这个值出现了52W多，这个有什么用吗？
  - dist1,dist2
    - dist1:训练集缺失60%，测试集缺失57%
dist2:训练集缺失95%，测试集缺失93%
数据缺失是否与isFraud有关呢？
dist1数据缺失的交易：其欺骗占比较大，未缺失的交易，其欺骗占比较小
dist2数据缺失的交易，其欺骗占比异常大（未欺骗/欺骗=9）
训练集未存在dist1,dist2数据都有的情况！
分析dist1,dist2同时缺失的情况？
dist1取值从0-1W多点，也是指数下降的分布，大部分数据都在较小的值，这应该是把数据按地域一块一块分类了
dist2取值同dist1
  - ProductCD
    - 训练集与测试集都无缺失值,
有W,C,R,H,C五种类型，不同类型isFraud的占比会不同，这个特征与isFraud的相关性应该挺高的
  - TransactionDT
    - 1. 从0开始15000000多（测试集从1.8千万到3.3千万，所以训练测试集数据时间未有重叠），
2. 分布的话基本是均匀分布，除了训练集刚刚开始那段时间数据较多，测试集快结束那段时间数据较多？
3. 这种整数数据类型该怎么分段？
4. 训练测试都未有缺失值
  - TransactionAmt
    - 1.数据极度右偏，有2行数据大于30000，其余数据都是1W以下，均值在135，中位数在69，
2. 观察amount<1000的分布，也是类似指数分布，不断下降的趋势。
3. 欺骗的平均amount在150，未欺骗的平均在135
4. 训练测试未有缺失值
  - P_emaildomain, R_emaildomain
    - P：9W多缺失值
R：45W多缺失值
都是一些邮箱域名
P与R分别代表什么意思呢？
R出现60种域名，P有59种域名
- train_identity
  - 144233条数据（不是所有交易数据都有对应的身份信息），除去transactionID(每行都唯一)，有40个特征
  - 37列特征存在缺失值
- 分支主题 3

XMind: ZEN 试用模式