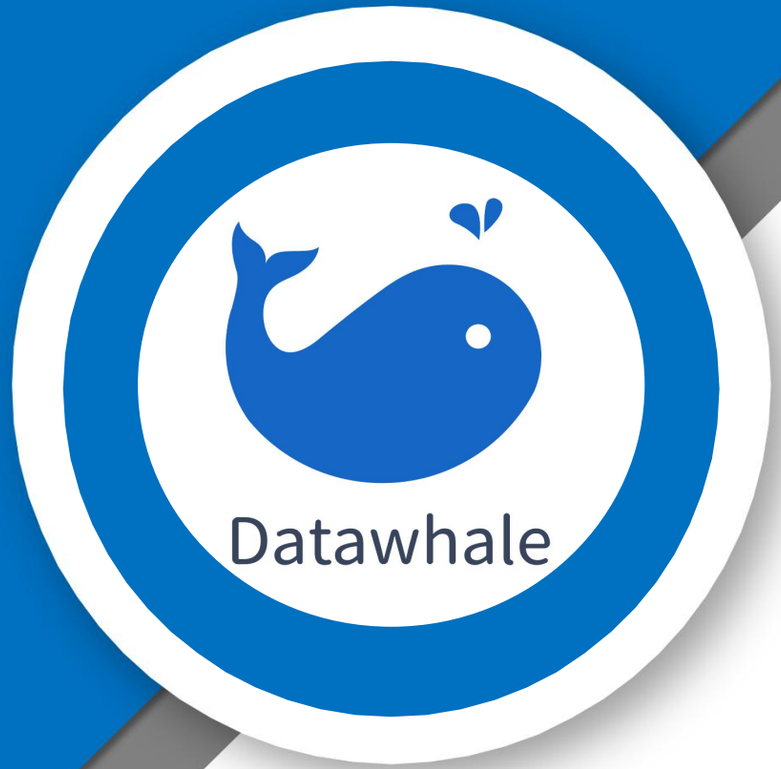

零基础入门金融风控

数据探索性分析和特征工程

言溪　　北京师范大学硕士

本次直播内容：

1. 了解赛题，产品，数据
2. 数据探索性分析
3. 特征工程
4. 模型
5. 问题解答

数据集和数据

数据初步了解：

读取数据集并了解数据集大小，原始特征维度（80，20，47，15）；

查看正负样本个数，判断样本是否均衡；

通过info熟悉数据类型（int64,float64,object）；

粗略查看数据集中各特征基本统计量；

缺失值和唯一值；

数据探索性分析--数据清洗

缺失值：

缺失值往往是不可避免的

不处理：少量样本缺失，不处理

补全：统计特征补全，手动补全，根据模型计算相似度补全

删除：大量样本缺失

异常值：

3segama法则（服从正太分布），箱型图

异常值检测

无效值：

对label无区分意义的特征删除

数据探索性分析--数据类型

类别型数据（定性）

- 有序
  - grade
- 无序
  - regionCode 地区编码

数值型数据（定量）

- 离散数值型数据
  - term 贷款期限
- 连续数值型数据
  - loanAmnt

离散性：

- 属性种类分布（自身）
- 不同种类的逾期率分布（和lebel的关系）
- 特征编码（数值型离散特征可不编码）
- 特征交叉

连续性：

- 方差（自身）
- 线性相关性（和lebel的关系）
- 分箱

数据探索性分析-深入了解数据

```
continuous_fea = ['loanAmnt','interestRate','installment','annualIncome','dti','openAcc',
                  'revolBal','revolUtil','totalAcc']
```

```
def get_statistical_charac(df,fea):
    res = {}
    res['fea_name'] = fea
    res['fea_mean'] = np.mean(train[fea])
    res['fea_median'] = np.median(train[fea])
    res['fea_max'] = max(train[fea])
    res['fea_min'] = min(train[fea])

    return res
```

```
for fea in continuous_fea:
    res = get_statistical_charac(train,fea)
    print(res)
```

```
{'fea_name': 'loanAmnt', 'fea_mean': 14416.818875, 'fea_median': 12000.0, 'fea_max': 40000.0, 'fea_min': 500.0}
{'fea_name': 'interestRate', 'fea_mean': 13.238391287551105, 'fea_median': 12.74, 'fea_max': 30.99, 'fea_min': 5.31}
{'fea_name': 'installment', 'fea_mean': 437.947723250128, 'fea_median': 375.135, 'fea_max': 1715.42, 'fea_min': 15.69}
{'fea_name': 'annualIncome', 'fea_mean': 76133.9104931876, 'fea_median': 65000.0, 'fea_max': 10999200.0, 'fea_min': 0.0}
{'fea_name': 'dti', 'fea_mean': 18.28455728648994, 'fea_median': nan, 'fea_max': 999.0, 'fea_min': -1.0}
{'fea_name': 'openAcc', 'fea_mean': 11.59802, 'fea_median': 11.0, 'fea_max': 86.0, 'fea_min': 0.0}
{'fea_name': 'revolBal', 'fea_mean': 16228.706505, 'fea_median': 11132.0, 'fea_max': 2904836.0, 'fea_min': 0.0}
{'fea_name': 'revolUtil', 'fea_mean': 51.790733812067835, 'fea_median': nan, 'fea_max': 892.3, 'fea_min': 0.0}
{'fea_name': 'totalAcc', 'fea_mean': 24.99886125, 'fea_median': 23.0, 'fea_max': 162.0, 'fea_min': 2.0}
```

```
#这个业务放款横跨了139天的时间间隔
len(train['issueDateDT'].unique())
```

```
139
```

```
len(train['interestRate'].unique())
#利率定价一般也是根据用户资质评分卡打的，或者就一些简单评分规则
```

```
641
```

term	all_sum	bad_sum	逾期率
3	606902	97126	0.160036
5	193098	62484	0.323587

grade	all_sum	bad_sum	逾期率
A	139661	8432	0.060375
B	233690	31079	0.132992
C	227118	51106	0.225020
D	119453	36296	0.303852
E	55661	21390	0.384291
F	19053	8641	0.453524
G	5364	2666	0.497017

- id 为贷款清单分配的唯一信用证标识
- loanAmnt 贷款金额
- term 贷款期限（year）
- interestRate 贷款利率
- installment 分期付款金额
- grade 贷款等级
- subGrade 贷款等级之子级
- employmentTitle 就业职称
- employmentLength 就业年限（年）
- homeOwnership 借款人在登记时提供的房屋所有权状况
- annualIncome 年收入
- verificationStatus 验证状态
- issueDate 贷款发放的月份
- purpose 借款人在贷款申请时的贷款用途类别
- postCode 借款人在贷款申请中提供的邮政编码的前3位数字
- regionCode 地区编码
- dti 债务收入比
- delinquency_2years 借款人过去2年信用档案中逾期30天以上的违约事件数
- ficoRangeLow 借款人在贷款发放时的fico所属的下限范围
- ficoRangeHigh 借款人在贷款发放时的fico所属的上限范围
- openAcc 借款人信用档案中未结信用额度的数量
- pubRec 贬损公共记录的数量
- pubRecBankruptcies 公开记录清除的数量
- revolBal 信贷周转余额合计
- revolUtil 循环额度利用率，或借款人使用的相对于所有可用循环信贷的信贷金额
- totalAcc 借款人信用档案中当前的信用额度总数
- initialListStatus 贷款的初始列表状态
- applicationType 表明贷款是个人申请还是与两个共同借款人的联合申请
- earliesCreditLine 借款人最早报告的信用额度开立的月份
- title 借款人提供的贷款名称
- policyCode 公开可用的策略_代码=1新产品不公开可用的策略_代码=2
- n系列匿名特征 匿名特征n0-n14，为一些贷款人行为计数特征的处理

特征工程

特征衍生：

一次衍生

- 时间特征（绝对时间转化相对时间）
- 非线性转化（log）
- 分箱

二次衍生

- 类别特征交叉
- 连续特征分箱再交叉

三次衍生

- 特征交叉后+统计量聚合

```
#转化成时间格式
for data in [train, testA]:
    data['issueDate'] = pd.to_datetime(data['issueDate'],format='%Y-%m-%d')
    startdate = datetime.datetime.strptime('2007-06-01', '%Y-%m-%d')
    #构造时间特征
    data['issueDateDT'] = data['issueDate'].apply(lambda x: x-startdate).dt.days
```

```
for data in [train, testA]:
    #贷款金额/分期付款金额 = 贷款期限
    data['loanTerm'] = data['loanAmnt'] / data['installment']
```

```
for feature in ['ficoRangeLow__applicationType', 'ficoRangeLow__ficoRangeHigh']:
    f1, f2 = feature.split('__')
    train[feature] = train[f1].astype(str) + '_' + train[f2].astype(str)
    testA[feature] = testA[f1].astype(str) + '_' + testA[f2].astype(str)
```

```
for grad in ['grade_homeOwnership', 'grade_term', 'grade_verificationStatus',
             'grade_purpose', 'grade_regionCode']:
    card1 = grad.split('_')[0]
    card2 = grad.split('_')[1]
    train[grad] = train[card1].astype(str) + '_' + train[card2].astype(str)
    testA[grad] = testA[card1].astype(str) + '_' + testA[card2].astype(str)

    train[grad + '_amt_mean'] = train[grad].map(
        (pd.concat([train[[grad, 'loanAmnt']], testA[[grad, 'loanAmnt']]], ignore_index=True)).groupby(
            [grad])['loanAmnt'].mean())
```

特征工程

特征分箱

- 目的：
  - 从模型效果上来看，特征分箱主要是为了降低变量的复杂性，减少变量噪音对模型的影响，提高自变量和因变量的相关度。从而使模型更加稳定。

- 数据分箱的对象：
  - 将连续变量离散化
  - 将多状态的离散变量合并成少状态

- 分箱的原因：
  - 数据的特征内的值跨度可能比较大（利用距离的算法模型）

- 分箱的优点：
  - 处理缺失值：当数据源可能存在缺失值，此时可以把null单独作为一个分箱。
  - 处理异常值：当数据中存在离群点时，可以把其通过分箱离散化处理，从而提高变量的鲁棒性（抗干扰能力）。例如，age若出现200这种异常值，可分入“age > 60”这个分箱里，排除影响。

- 分箱的基本原则：
  - 最小分箱占比不低于5%
  - 箱内不能全部是好客户
  - 连续箱单调

- 分箱的方法：
  - 根据log等非线性函数映射分箱，卡方分箱，决策树分箱。

特征工程

WOE和IV：

评分卡建模流程中，WOE（Weight of Evidence）常用于特征变换，IV（Information Value）则用来衡量特征的预测能力。

为什么woe不能直接作为衡量变量预测能力的指标呢?

loanAmnt,interestRate,installment,dti,openAcc,revolUtil和逾期率正比关系
annualIncome,totalAcc和逾期率反比关系
revolBal 倒u结构

```
{'loanAmnt_iv': 0.038276098495837 69,
 'interestRate_iv': 0.47021950983989613,
 'installment_iv': 0.040730489480097 66,
 'annualIncome_iv': 0.030151238516772438,
 'dti_iv': 0.072720261973468 38,
 'openAcc_iv': 0.004616785403783 3704,
 'revolBal_iv': 0.004455092339434142,
 'revolUtil_iv': 0.025072879341782518,
 'totalAcc_iv': 0.002076623819504 4427}
```

特征工程

特征筛选：

特征选择技术可以精简掉无用的特征，以降低最终模型的复杂性，它的最终目的是得到一个简约模型，在不降低预测准确率或对预测准确率影响不大的情况下提高计算速度。特征选择不是为了减少训练时间（实际上，一些技术会增加总体训练时间），而是为了减少模型评分时间。

特征选择的方法：

- Filter（过滤）
  - 方差选择法
  - 相关系数法（pearson 相关系数）
  - 卡方检验
  - 互信息法
- Wrapper （封装）
  - 递归特征消除法
- Embedded（嵌入）
  - 基于惩罚项的特征选择法
  - 基于树模型的特征选择

特征工程

特征编码：

- 频次编码
- ont-hot编码
- labelencode编码
- WOE编码

模型融合

模型：

- xgb-lgb-catboost
- woe编码后用logistic
- nn

解答

猜变量：homeOwnership 各类别中文含义代表什么？

借款人在登记时提供的房屋所有权状况：

1、初始登记： 新建房屋竣工后，权利人申请初始登记；

2、转移登记： 房屋买卖、交换、赠与、继承、调拨、以房地产作价入股或者作为合作条件与他人成立法人或者其他组织、法人或者其他组织合并（分立）、以房地产清偿债务、以其他合法方式使房屋权属发生变更的，当事人申请转移登记；

3、变更登记：房屋所有权人（共有权人）名称改变、房屋坐落的地址变更、房屋面积增加或减少、房屋登记状况变更的，权利人申请变更登记；

4、他项权利（抵押权）登记：设定房屋抵押权，当事人申请房屋抵押登记。登记后抵押情况发生变更的，当事人申请抵押变更登记；

5、注销登记： 房屋灭失、抵押权终止、房屋权利灭失的，权利人申请注销登记；

6、补证、换证登记：证书（证明）遗失的，申请补证登记；证书（证明）破损的，申请换证登记。

```
train['homeOwnership'].value_counts()
```

```
0    395732
1    317660
2     86309
3       185
5        81
4        33
Name: homeOwnership, dtype: int64
```

```
grade_risk = train.groupby('homeOwnership')['isDefault'].agg(all_sum = np.size, bad_sum = np.sum)
grade_risk['逾期率'] = grade_risk['bad_sum']/grade_risk['all_sum']
```

```
grade_risk
```

homeOwnership	all_sum	bad_sum	逾期率
0	395732	67882	0.171535
1	317660	73731	0.232107
2	86309	17935	0.207800
3	185	38	0.205405
4	33	5	0.151515
5	81	19	0.234568