法律声明

- 本课件包括：演示文稿，示例，代码，题库，视频和声音等，小象学院拥有完全知识产权的权利；只限于善意学习者在本课程使用，不得在课程范围外向任何第三方散播。任何其他人或机构不得盗版、复制、仿造其中的创意，我们将保留一切通过法律手段追究违反者的权利。

- 课程详情请咨询
  - 微信公众号：大数据分析挖掘
  - 新浪微博：ChinaHadoop

流失预警模型中的数据预处理和特征衍生

目录

极端值的处理

极端值的定义

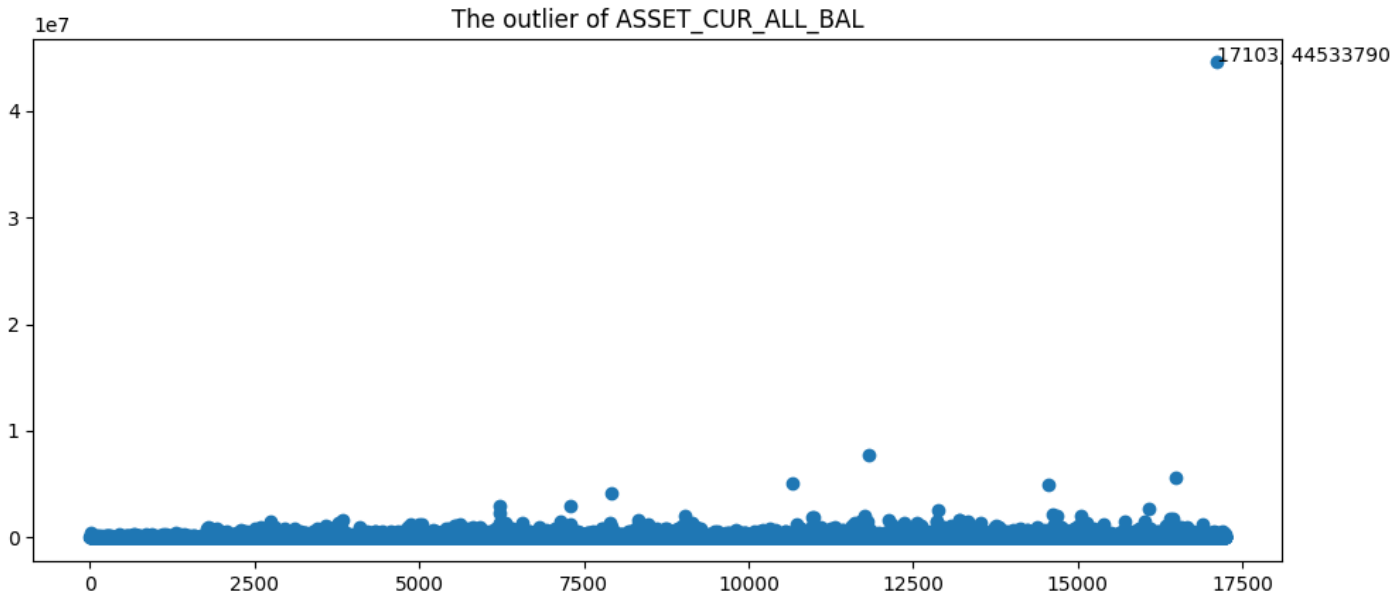
又称离群值，往往会扭曲预测结果并影响模型精度。回归模型中离群值的影响尤其大，使用该模型时我们需要对其进行检测和处理。

极端值检测的重要性

- 处理离群值或者极端值并不是数据建模的必要流程，然而，了解它们对预测模型的影响也是大有裨益的。
- 数据分析师们需要自己判断处理离群值的必要性，并结合实际问题选取处理方法。
- 检测离群值的重要性：由于离群值的存在，模型的估计和预测可能会有很大的偏差或者变化
- 可以选择对极端值不敏感的模型，例如KNN，决策树

极端值的处理

极端值的可视化检验

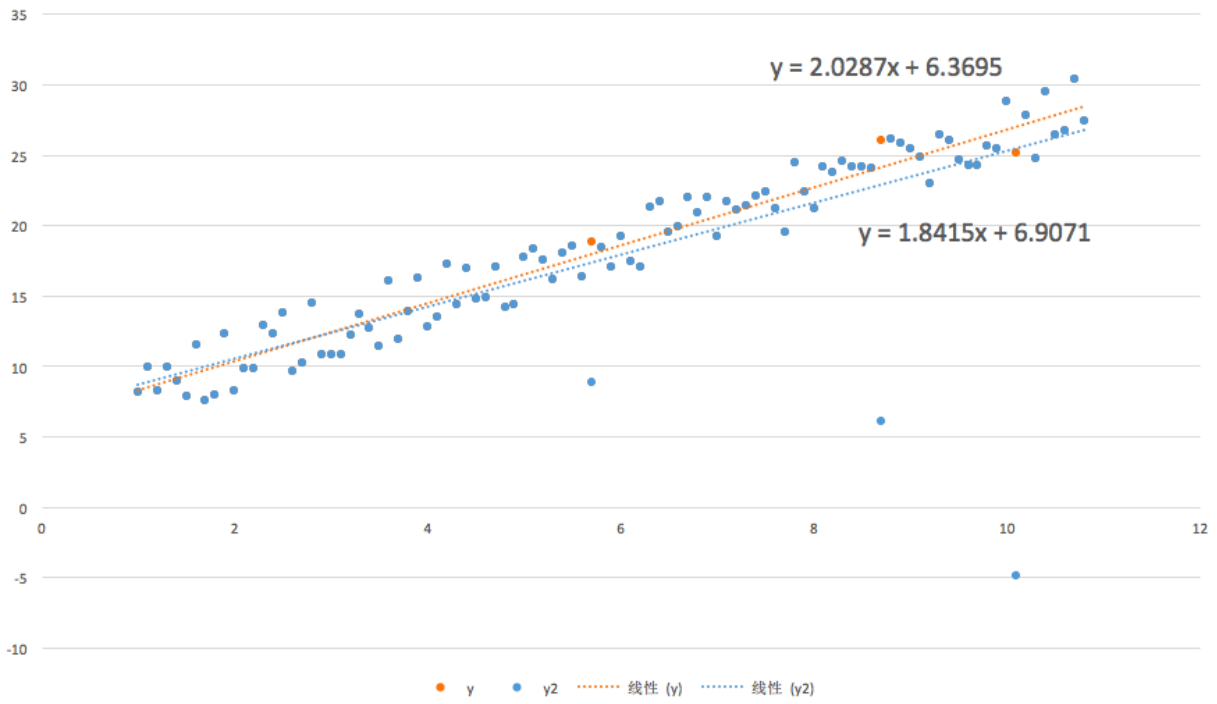


- 大于10,000,000的样本只占0.006%
- 除去极端值外，其余样本的均值是115,443，标准差是179,359
- 极端值是均值的386倍，偏离247个标准差

极端值的处理

极端值对模型的影响

通常情况下，极端值对模型会带来一定的偏差，例如在线性回归中，极端值会显著地影响模型的参数估计。



极端值的处理

极端值的处理

- 人为降低极端值到某个正常的值，例如用95%的分位点代替

  例：因为透支的原因信用卡使用额度超过100%，可以用100%来代替

- 删除极端值

  例：极个别持卡人的年龄超过85岁

- 单独建模型

  例：信用卡额度特别高

目录

缺失值的处理

种类

- 完全随机缺失：缺失值跟其他变量无关，例如婚姻状况的缺失
- 随机缺失：缺失值依赖于其他变量，例如“配偶姓名”的缺失取决于“婚姻状况”
- 完全非随机缺失：缺失值依赖于自己，例如高收入人群不愿易提供家庭收入

处理方法

- 删除有缺失值的属性或者样本(土豪行为)
- 插补填充(常用于完全随机缺失且缺失度不高的情形中)
- 将缺失当成一种属性值(常用于完全非随机缺失)

缺失值的处理

连续变量缺失值的处理

对于完全随机缺失，当缺失率不高时，可以：

- 用常数补缺，例如均值

  特别地，如果存在极端值，要考虑是否剔除极端值后再计算均值

- 从非缺失值中随机抽样赋予缺失样本

对于依赖于其他某变量的随机缺失，可以在同一层内，用完全随机缺失的方法进行补缺

- 例如：变量“收入”取决于“工作状态”。当“工作状态”＝“有工作”时，缺失的“收入”可以用所有“有工作”的持卡人的已知收入的均值代替

对于完全非随机缺失，可以当成一种属性，将该变量转化成类别变量

缺失值的处理

类别变量缺失值的处理

当缺失率很低时

- 最常出现的类别补缺
- 可以从其他已知的样本中随机抽样进行补缺

当缺失率很高时

- 考虑剔除该属性

当缺失率介于“很低”和“很高”时

- 可以当成一种类别

目录

特殊变量的处理

类别变量

- 表述类目的变量，通常没有“次序”的概念，且取值范围有限
  - 性别，行业，信用卡种类
- 有些模型可以直接读如类别变量
  - 决策树
- 有些模型不能直接读如类别变量
  - 回归模型
  - 神经网络
  - 有“距离”度量的模型(SVM，kNN等)

    (注：计算距离前需要归一化)

特殊变量的处理

类别变量(续)

类别变量不能直接放入模型时，需要编码:以数值的形式代替原有值

- One – hot编码
- Dummy
- 浓度编码
- WOE编码

特殊变量的处理

日期/时间型变量

- 常常以字符串的形式出现，例如“2017-04-01 12:00:05”
- 本质上是数值型

  Excel:1900-01-01为第一天
- 可以基于某个基准日期，转化为天数

  以观察点为基准，将所有开户日期转为距离观察点的天数(month-on-book)

目录

构建流失行为的特征

内部自有数据

- 丰富的内部交易明细数据，包括本币活期储蓄波动率，本币活期储蓄月日均余额，。。。。，电话银行总交易笔数
- 可以构建的特征：
  - 不同交易的数额的比例
  - 单笔交易的平均数额
  - 某种交易的笔数占全部交易笔数的比例

构建流失行为的特征

内部自有数据(续)

例如

$$\text{活期日均余额比率} = \frac{\text{本币活期储蓄月日均余额}}{\text{本币活期储蓄月日均余额} + \text{本币定期月日均余额}}$$

最大波动率

$= \max\{\text{本币一年以下波动，本币一年以上波动率，储蓄类资产波动率，本币储蓄波动率}\}$

$$\text{本币活期续存交易金额} = \frac{\text{本币活期续存交易金额}}{\text{本币活期续存交易笔数}}$$

$$\text{本币平均活期转账交易金额} = \frac{\text{本币活期转账交易金额}}{\text{本币活期转账交易笔数}}$$

构建流失行为的特征

□ 内部自有数据(续)

信息存在冗余，需要按情况进行剔除

情况一

“本币活期月日均余额占比”＝1－“本币定期月日均余额占比”

变量“本币活期月日均余额占比”与“本币定期月日均余额占比”存在冗余性，知道其一必知道其二，需要剔除一个

情况二

“资产当前总余额”＝“本币储蓄当前总余额”＋“外币储蓄当前总余额”

如果是(广义)线性回归模型，三者不能同时放进模型。对于树模型，可以将其中任意两个放进模型，剩余的做转换

构建流失行为的特征

外部数据包含了客户在电信运营商的详情

包括：

- 通话时间与次数
- 话费详情
- 特定的呼叫行为
- 其他信息

可以衍生的特征

月平均通话时间的变化 = 过去三个月月平均通话时间 − 过去六个月月平均通话时间
月平均通话次数的变化 = 过去三个月月平均通话次数 − 过去六个月月平均通话次数
月平均缴纳话费的变化 = 过去三个月月平均缴纳话费 − 过去六个月月平均缴纳话费

疑问

- 问题答疑： http://www.xxwenda.com/
  - 可邀请老师或者其他人回答问题

联系我们

小象学院：互联网新技术在线教育领航者

- 微信公众号：大数据分析挖掘
- 新浪微博：ChinaHadoop