Concat: github: lambda\_xmu

- CATALOG

本文在包大人基础之上进行补充: PPT: Kaggle 比赛的进阶技巧和国内比赛前十套路 视频: https://www.bilibili.com/video/av57480953/?p=2

## 特征工程

#### 编码角度

#### 类别特征:

- 频度统计 count:
  - 。 优势: 可以解决长尾问题,将出现次数少的进行合并
- 转化率 target encoding,即 ctr:
  - 稀疏类别和类别较少时不宜做 target encoding
- one-hot
- Label Encoding
- Frequency Encoding
- leave one out • WOE & IV
- embedding

#### 连续性特征

- 转为类别值,使用类别特征
- min
- max
- mean standard
- quantile
- 分箱

#### 组合特征

- 对象
  - 。 类别+类别=更细类别
  - 。 类别+连续=原类别
  - 。 连续+连续=新连续
- 操作
  - o sum
  - difference
  - product
  - quotient

## 时间序列特征

- 时间窗+统计(min, max, mean, median, std)
  - 。 刻画这一个时间窗内的信息
- 特殊时间
  - 。 指示变量

# 图特征

- pagerank
- graph embedding

## 其他

- 降维
- 聚类

## 业务角度

- 反欺诈
  - 。 设备唯一性
  - 。 行为密度(短时间内操作多少次)
  - 。 行为平稳性(是否经常换个人信息)
- 二手售卖可能性
  - 。 出价合理度: (出价-同类出售均值)/同类出售均值
- 鼠标滑动验证码 。 加速度
  - 。 减速度

## 特征选择

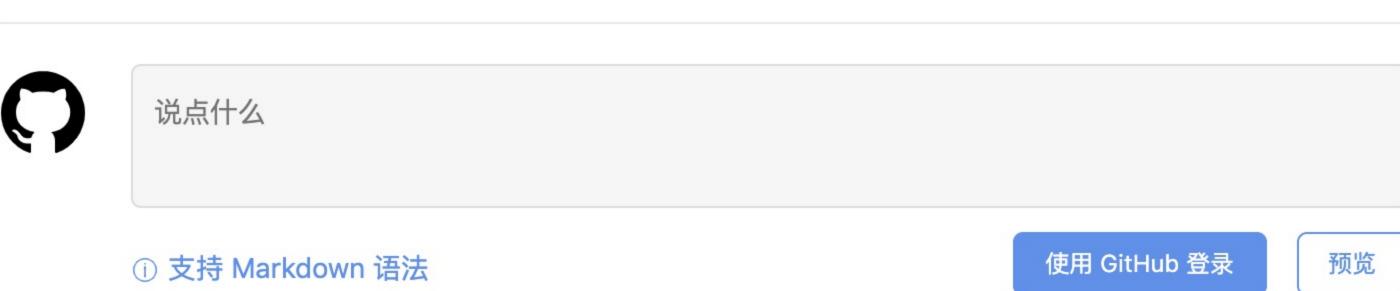
- 模型选择
- 。 重要性排序 • 统计指标

。 相关性

**PREVIOUS** 2019-08-27-2019CCF-CAR-SALES-EDA

**NEXT** 2019-09-04-2019CCF-WORK-PIECE-EDA-PART2

0条评论 未登录用户 ~



来做第一个留言的人吧!

**FEATURED TAGS Data Competition** EDA Feature Engineering Processing Baseline

**FRIENDS** 

drop-out Smile



