# 特征工程

## 基本信息

数据完整性较好, 无空缺值

数值型变量: hour,

类别型变量: 除了hour

## Onehot\_A\_hour

对 hour 进行处理, 生成如下新特征, 并进行onehot处理

|  |  |
| --- | --- |
| 新特征 | 含义 |
| 'hour' | 小时 |
| 'week' | 星期 |
| 'workingday' | 是否工作日 |
| 'timequantum' | 时间段(0-23,每5个小时为一段) |

## Onehot\_A\_his

定义新特征: 用户, 为device\_id,

但是device\_id=’a99f214a’出现的频率非常大, 则当device\_id=’a99f214a’,

用户为device ip + device model

对每个用户历史浏览做统计, 按天为单位进行次数统计.

分别统计点击次数和浏览次数,

由于猜测C17为广告的类型分类, 也对C17进行统计, 统计结果可以看做用户感兴趣的广告.

统计最感兴趣, 和第二感兴趣的广告的类别

|  |  |
| --- | --- |
| 特征名 | 含义 |
| click\_times | 点击次数 |
| total\_times | 浏览次数 |
| C17\_Fir | 最感兴趣广告类别 |
| C17\_Sec | 第二感兴趣广告类别 |

## Onehot\_A\_cat

对每个类别特征的每个取值计数, 并只对该特征中频率>=10的类别名进行onehot, <10的特征值, 统一设为字符串'less\_threshold'

## Onehot\_B\_cat

先对特征进行合并, 再以Onehot\_A的方法处理.

合并方法:不同特征的不同值, 先转为字符串, 再以分隔符'&'相连接, 如下表所示

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 原始特征 | 总类别数 | -> | 新特征 | 总类别数 |
| 'C1','banner\_pos' | 7+7 | -> | 'C1BP' | 16 |
| 'site\_id','site\_domain','site\_category' | 4737+7745+26 | -> | 'site' | 21637 |
| 'app\_id','app\_domain','app\_category' | 8552+559+36 | -> | 'app' | 8558 |
| 'device\_id','device\_ip' | 2686408+6729486 | -> | 'dev\_idip' | 8963496 |
| 'device\_model', 'device\_type','device\_conn\_type' | 8251+5+4 | -> | 'dev\_types' | 16069 |
| 'C14','C17' | 2626+435 | -> | 'C1417' | 2626 |
| 'C15','C16' | 8+9 | -> | 'C1516' | 10 |
| 'C18','C20' | 4+172 | -> | 'C1820' | 636 |
| 'C19','C21' | 68+60 | -> | 'C1921' | 220 |

## Onehot\_xgb

对Onehot\_A\_cat和Onehot\_B\_cat 进行树模型处理, 并对生成的xgboost模型的叶子结点的索引输出进行onehot编码

## Onehot\_ffm

1、hour，提取属于哪个小时，转换成0~23取值类别变量

2、device\_id，分成两类，等于‘a99f214a‘与其他

3、site\_id、site\_domain、app\_id、device\_model处理方式：计算每个类别的出现频次，低于20的归类为其他，高于20保持不变

4、C15 & C16，两列组合成一列（可能代表设备分辨率）

5、device\_ip，统计每个site\_id和app\_id的用户数，取log10并取整，生产两个新特征两个新特征代表每个site和app的用户规模。（当做类别变量）

6、其余特征统一作为类别特征