# 数据探索

## 数据量

|  |  |  |
| --- | --- | --- |
| 名称 | 大小 | 样本数 |
| train.csv | 6.3G | 40M(10天) |
| test.csv | 0.7G | 4M(1天) |

## 特征含义

|  |  |
| --- | --- |
| 名称 | 含义 |
| id | ad identifier |
| click | 1是点击 0/1 for non-click/click |
| hour | 时间, YYMMDDHH, so 14091123 means 2014年09月11号23:00 UTC. |
| C1 | 广告大小 |
| banner\_pos | 广告位置 |
| site\_id | 网站ID号 |
| site\_domain | 网站领域 |
| site\_category | 网站类别 |
| app\_id | appID号 |
| app\_domain | 应用领域 |
| app\_category | 应用类别 |
| device\_id | 设备ID号 |
| device\_ip | 设备ip地址 |
| device\_model | 设备型号, 如iphone5/iphone4 |
| device\_type | 设备类型, 如智能手机/平板电脑 |
| device\_conn\_type | 连接设备类型 |
| C14-C21 | 未知分类变量 anonymized categorical variables |
| C14 | 广告ID |
| C15 | 设备分辨率 |
| C16 | 设备分辨率 |
| C17 | 广告类型 |
| C18 |  |
| C19 |  |
| C20 |  |
| C21 |  |

## 特征基本信息

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 特征 | 数据类型 | 类别数 | 频率最大类别名 | 占总样本比例(%) | 空缺值数 |
| id | -- | -- | -- | -- | -- |
| click | int64 | 2 | 0 | 83 | 0 |
| hour | int64 | 240 | 14102209 | 1.1 | 0 |
| C1 | int64 | 7 | 1005 | 91.9 | 0 |
| banner\_pos | int64 | 7 | 0 | 72.0 | 0 |
| site\_id | object | 4737 | '85f751fd' | 36.1 | 0 |
| site\_domain | object | 7745 | 'c4e18dd6' | 37.4 | 0 |
| site\_category | object | 26 | '50e219e0' | 40.9 | 0 |
| app\_id | object | 8552 | 'ecad2386' | 63.9 | 0 |
| app\_domain | object | 559 | '7801e8d9' | 67.4 | 0 |
| app\_category | object | 36 | '07d7df22' | 64.7 | 0 |
| device\_id | object | 2686408 | 'a99f214a' | 82.5 | 0 |
| device\_ip | object | 6729486 | '6b9769f2' | 0.5 | 0 |
| device\_model | object | 8251 | '8a4875bd' | 6.1 | 0 |
| device\_type | int64 | 5 | 1 | 92.3 | 0 |
| device\_conn\_type | int64 | 4 | 0 | 86.3 | 0 |
| C14 | int64 | 2626 | 4687 | 2.3 | 0 |
| C15 | int64 | 8 | 320 | 93.3 | 0 |
| C16 | int64 | 9 | 50 | 94.3 | 0 |
| C17 | int64 | 435 | 1722 | 11.2 | 0 |
| C18 | int64 | 4 | 0 | 41.9 | 0 |
| C19 | int64 | 68 | 35 | 30.1 | 0 |
| C20 | int64 | 172 | -1 | 46.8 | 0 |
| C21 | int64 | 60 | 23 | 22.0 | 0 |

## Click

Click = 0有33.6M个样本, Click = 1有6.8M个样本,

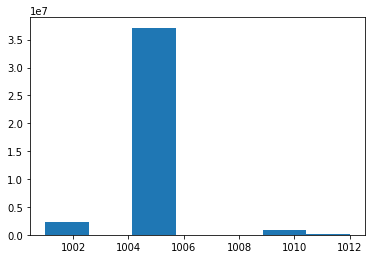


## Hour

一共10天, 每天24小时, 共240个类别

## C1

未知分类变量, 有7个类别, 由于和banner\_pos相关性很强, 猜测为广告展示的面积



## banner\_pos

广告展示的位置, 有7个类别

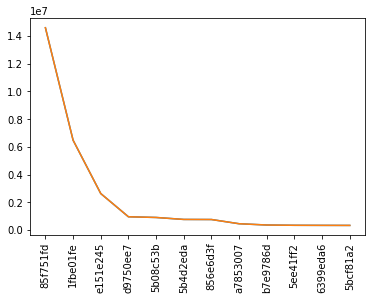


## site\_id

网站ID号, 有4737个类别, 频率排行前5的类别如下表所示

|  |  |
| --- | --- |
| 类别名 | 频率 |
| 85f751fd | 14596137 |
| 1fbe01fe | 6486150 |
| e151e245 | 2637747 |
| d9750ee7 | 963745 |
| 5b08c53b | 913325 |

对出现次数最多的前12个分类进行可视化

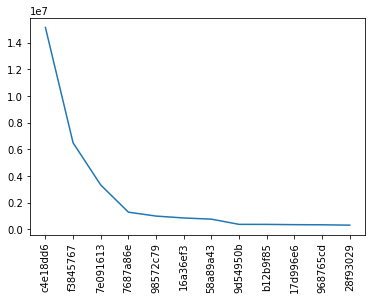


## site\_domain

网站领域, 有7745个类别, 频率排行前5的类别如下表所示

|  |  |
| --- | --- |
| 类别名 | 频率 |
| c4e18dd6 | 15131739 |
| f3845767 | 6486150 |
| 7e091613 | 3325008 |
| 7687a86e | 1290165 |
| 98572c79 | 996816 |

对出现次数最多的前12个分类进行可视化



## site\_category

网站领域, 有26个类别, 频率排行前5的类别如下表所示

|  |  |
| --- | --- |
| 类别名 | 频率 |
| 50e219e0 | 16537234 |
| f028772b | 12657073 |
| 28905ebd | 7377208 |
| 3e814130 | 3050306 |
| f66779e6 | 252451 |

对出现次数最多的前12个分类进行可视化