**大数据分析与处理**

**课程项目选题说明文档**

**教师：尚家兴**

**助教：管梦雅**

**目 录**

[1 数据分析-学术前沿趋势分析 1](#_Toc99896469)

[1.1 赛题描述 1](#_Toc99896470)

[1.2 数据说明 1](#_Toc99896471)

[1.3 读取数据 2](#_Toc99896472)

[1.4 任务描述 2](#_Toc99896473)

[1.5 赛题网址 2](#_Toc99896474)

[2 数据分析 -汽车产品聚类分析 2](#_Toc99896475)

[2.1 赛题描述 2](#_Toc99896476)

[2.2 数据描述 2](#_Toc99896477)

[2.3 赛题任务 4](#_Toc99896478)

[2.4 赛题网址 4](#_Toc99896479)

[3 网格案件潮汐预测 4](#_Toc99896480)

[3.1 赛题任务 4](#_Toc99896481)

[3.2 数据描述 4](#_Toc99896482)

[3.3 评价指标 5](#_Toc99896483)

[3.4 赛题网址 5](#_Toc99896484)

[4 数据挖掘 - 二手车交易价格预测 5](#_Toc99896485)

[4.1 赛题描述 5](#_Toc99896486)

[4.2 数据描述 5](#_Toc99896487)

[4.3 评价指标 6](#_Toc99896488)

[4.4 赛题网址 6](#_Toc99896489)

[5 新浪微博互动预测 6](#_Toc99896490)

[5.1 赛题描述 6](#_Toc99896491)

[5.2 数据描述 7](#_Toc99896492)

[5.3 评价指标 7](#_Toc99896493)

[5.4 赛题网址 8](#_Toc99896494)

[6 心跳信号分类预测 8](#_Toc99896495)

[6.1 赛题描述 8](#_Toc99896496)

[6.2 数据描述 9](#_Toc99896497)

[6.3 评价指标 9](#_Toc99896498)

[6.4 赛题网址 9](#_Toc99896499)

[7 一见钟情指数预测 9](#_Toc99896500)

[7.1 赛题描述 9](#_Toc99896501)

[7.2 数据描述 10](#_Toc99896502)

[7.3 赛题任务 10](#_Toc99896503)

[7.4 赛题网址 10](#_Toc99896504)

[8 贷款违约预测 10](#_Toc99896505)

[8.1 赛题描述 10](#_Toc99896506)

[8.2 数据描述 10](#_Toc99896507)

[8.3 评价指标 12](#_Toc99896508)

[8.4 赛题网址 12](#_Toc99896509)

[9 挖掘幸福感 12](#_Toc99896510)

[9.1 赛题背景 12](#_Toc99896511)

[9.2 赛题说明 12](#_Toc99896512)

[9.3 数据描述 13](#_Toc99896513)

[9.4 评测指标 13](#_Toc99896514)

[9.5 赛题网址 13](#_Toc99896515)

[10 安全恶意程序检测 13](#_Toc99896516)

[10.1 赛题背景 13](#_Toc99896517)

[10.2 赛题说明 13](#_Toc99896518)

[10.3 数据描述 14](#_Toc99896519)

[10.4 评测指标 14](#_Toc99896520)

[10.5 赛题网址 15](#_Toc99896521)

# 数据分析-学术前沿趋势分析 复杂社交网络❌|无长期赛

## 1.1 赛题描述

arXiv 是一个重要的学术公开网站，也是搜索、浏览和下载学术论文的重要工具。arXiv论文涵盖的范围非常广，涉及物理学的庞大分支和计算机科学的众多子学科，如数学、统计学、电气工程、定量生物学和经济学等等。

本次赛题将使用arXiv在公开的论文数据集，赛题任务为通过数据分析能够挖掘出最近学术的发展趋势和学术关键词。

## 1.2 数据说明

数据存放在“arxiv-metadata-oai-2019.json.zip”中，其字段表如表1所示，

| **Field** | **Description** |
| --- | --- |
| id | arXiv ID，可用于访问论文 |
| submitter | 论文提交者 |
| authors | 论文作者 |
| title | 论文标题 |
| comments | 论文页数和图表等其他信息 |
| journal-ref | 论文发表的期刊的信息 |
| doi | 数字对象标识符，https://www.doi.org |
| report-no | 报告编号 |
| categories | 论文在 arXiv 系统的所属类别或标签 |
| license | 文章的许可证 |
| abstract | 论文摘要 |
| versions | 论文版本 |
| authors\_parsed | 作者信息（处理后） |

表1 赛题1字段表

## 1.3 读取数据

data = [] #初始化

#使用with语句优势：1.自动关闭文件句柄；2.自动显示（处理）文件读取数据异常

with open("arxiv-metadata-oai-snapshot.json", r) as f:

for line in f:

data.append(json.loads(line))

data = pd.DataFrame(data) #将list变为dataframe格式，方便使用pandas进行分析

data.shape #显示数据大小

## 1.4 任务描述

任务1：论文数量统计（数据统计任务）：统计2019年全年，计算机各个方向论文数量；

任务2：论文作者统计（数据统计任务）：统计所有论文作者出现评率Top10的姓名；

任务3：论文代码统计（数据统计任务）：统计所有论文类别下包含源代码论文的比例；

任务4：论文分类（数据建模任务）：利用已有数据建模，对新论文进行类别分类；

任务5：作者关联（数据建模任务）：对论文作者关系进行建模，统计最常出现的作者关系；

## 1.5 赛题网址

<https://tianchi.aliyun.com/competition/entrance/531866/information>

# 数据分析 -汽车产品聚类分析 聚类|无长期赛

## 2.1 赛题描述

赛题以竞品分析为背景，通过数据的聚类，为汽车提供聚类分类。对于指定的车型，可以通过聚类分析找到其竞品车型。通过这道赛题，鼓励学习者利用车型数据，进行车型画像的分析，为产品的定位，竞品分析提供数据决策。

## 2.2 数据描述

数据源：car\_price.csv，数据包括了205款车的26个字段。

|  |  |  |
| --- | --- | --- |
| 序号 | 字段 | 说明 |
| 1 | Car\_ID | Unique id of each observation (Interger) |
| 2 | Symboling | Its assigned insurance risk rating, A value of +3 indicates that the auto is risky, -3 that it is probably pretty safe.(Categorical) |
| 3 | carCompany | Name of car company (Categorical) |
| 4 | fueltype | Car fuel type i.e gas or diesel (Categorical) |
| 5 | aspiration | Aspiration used in a car (Categorical) |
| 6 | doornumber | Number of doors in a car (Categorical) |
| 7 | carbody | body of car (Categorical) |
| 8 | drivewheel | type of drive wheel (Categorical) |
| 9 | enginelocation | Location of car engine (Categorical) |
| 10 | wheelbase | Weelbase of car (Numeric) |
| 11 | carlength | Length of car (Numeric) |
| 12 | carwidth | Width of car (Numeric) |
| 13 | carheight | height of car (Numeric) |
| 14 | curbweight | The weight of a car without occupants or baggage. (Numeric) |
| 15 | enginetype | Type of engine. (Categorical) |
| 16 | cylindernumber | cylinder placed in the car (Categorical) |
| 17 | enginesize | Size of car (Numeric) |
| 18 | fuelsystem | Fuel system of car (Categorical) |
| 19 | boreratio | Boreratio of car (Numeric) |
| 20 | stroke | Stroke or volume inside the engine (Numeric) |
| 21 | compressionratio | compression ratio of car (Numeric) |
| 22 | horsepower | Horsepower (Numeric) |
| 23 | peakrpm | car peak rpm (Numeric) |
| 24 | citympg | Mileage in city (Numeric) |
| 25 | highwaympg | Mileage on highway (Numeric) |
| 26 | price(Dependent variable) | Price of car (Numeric) |

表2 赛题2字段表

## 2.3 赛题任务

对该汽车数据进行聚类分析，并找到vokswagen汽车的相应竞品。（聚类分析是常用的数据分析方法之一，不仅可以帮助我们对用户进行分组，还可以帮我们对产品进行分组（比如竞品分析） 这里的聚类个数选手可以根据数据集的特点自己指定，并说明聚类的依据）

## 2.4 赛题网址

<https://tianchi.aliyun.com/competition/entrance/531892/information>

# 网格案件潮汐预测 网格预测×|无长期赛

## 3.1 赛题任务

从真实场景和实际应用出发，根据历史三年（2017年1月-2019年12月）各个网格各类案件的时空数据，需要参赛者建立准确的预测模型，来预测未来各个网格各类案件的数量。要求：预测2020年1月-2020年12月各个网格各类案件的数量。

## 3.2 数据描述

提供历史三年所有案件各个维度数据，包括时空信息、案件类型等。选手可以下载数据，在本地进行算法调试，在比赛页面提交结果。

数据样本如下：

| **字段名** | **类型说明** |
| --- | --- |
| **案件编号** | String |
| **案件发生时间**  **String** | String |
| **责任网格** | String 共32个网格 |
| **案件类型** | String 共6种案件类型 |

表3 赛题3字段表

备注：训练集需要选手自己统计，每个月，每个责任网格，每类案件的数量作为训练集的标签（caseload）。

## 3.3 评价指标

赛题选用的评价指标为测试集的MSE。

## 3.4 赛题网址

[http://data.sd.gov.cn/cmpt/cmptDetail.html?id=52](https://tianchi.aliyun.com/competition/entrance/531810/information)

# 数据挖掘 - 二手车交易价格预测 预测类|长期赛，15265

## 4.1 赛题描述

赛题以预测二手车的交易价格为任务，该数据来自某交易平台的二手车交易记录。

## 4.2 数据描述

该数据集总数据量超过40w，包含31列变量信息，其中15列为匿名变量。为了保证比赛的公平性，将会从中抽取15万条作为训练集，5万条作为测试集A，5万条作为测试集B，同时会对name、model、brand和regionCode等信息进行脱敏。数据集字段表如表5所示：

| **Field** | **Description** |
| --- | --- |
| SaleID | 交易ID，唯一编码 |
| name | 汽车交易名称，已脱敏 |
| regDate | 汽车注册日期，例如20160101，2016年01月01日 |
| model | 车型编码，已脱敏 |
| brand | 汽车品牌，已脱敏 |
| bodyType | 车身类型：豪华轿车：0，微型车：1，厢型车：2，大巴车：3，敞篷车：4，双门汽车：5，商务车：6，搅拌车：7 |
| fuelType | 燃油类型：汽油：0，柴油：1，液化石油气：2，天然气：3，混合动力：4，其他：5，电动：6 |
| gearbox | 变速箱：手动：0，自动：1 |
| power | 发动机功率：范围 [ 0, 600 ] |
| kilometer | 汽车已行驶公里，单位万km |
| notRepairedDamage | 汽车有尚未修复的损坏：是：0，否：1 |
| regionCode | 地区编码，已脱敏 |
| seller | 销售方：个体：0，非个体：1 |
| offerType | 报价类型：提供：0，请求：1 |
| creatDate | 汽车上线时间，即开始售卖时间 |
| price | 二手车交易价格（预测目标） |
| v系列特征 | 匿名特征，包含v0-14在内15个匿名特征 |

表4 赛题4字段表

## 4.3 评价指标

评价标准为MAE(Mean Absolute Error)，如公式(1)所示，

 (1)

其中，表示编号为的二手车的真实价格，表示编号为的二手车的预测价格。

## 4.4 赛题网址

<https://tianchi.aliyun.com/competition/entrance/231784/information>

# 新浪微博互动预测 | 长期赛，5532，给了Baseline，提升空间大 ×

## 5.1 赛题描述

对于一条原创博文而言,转发、评论、赞等互动行为能够体现出用户对于博文内容的兴趣程度，也是对博文进行分发控制的重要参考指标。本届赛题的任务就是根据抽样用户的原创博文在发表一天后的转发、评论、赞总数，建立博文的互动模型，并预测用户后续博文在发表一天后的互动情况。

## 5.2 数据描述

1. 训练数据（weibo\_train\_data(new)）2015-02-01至2015-07-31

博文的全部信息都映射为一行数据。其中对用户做了一定抽样，获取了抽样用户半年的原创博文，对用户标记和博文标记做了加密,发博时间精确到天级别。字段表如表6所示，

|  |  |  |
| --- | --- | --- |
| **Field** | **Field Description** | **Extraction Description** |
| uid | 用户标记 | 抽样&字段加密 |
| mid | 博文标记 | 抽样&字段加密 |
| time | 发博时间 | 精确到天 |
| forward\_count | 博文发表一周后的转发数 |  |
| comment\_count | 博文发表一周后的评论数 |  |
| like\_count | 博文发表一周后的赞数 |  |
| content | 博文内容 |  |

表5 赛题5训练数据字段表

1. 预测数据（weibo\_predict\_data(new)）2015-08-01至2015-08-31

预测数据表示等待预测的用户博文，字段表如表7所示，

|  |  |  |
| --- | --- | --- |
| 字段 | 字段说明 | 提取说明 |
| uid | 用户标记 | 抽样&字段加密 |
| mid | 博文标记 | 抽样&字段加密 |
| time | 发博时间 | 精确到天 |
| content | 博文内容 |  |

表6 赛题5预测数据字段表

1. 结果数据（weibo\_result\_data）

赛题任务为对预测数据（weibo\_predict\_data）中每条博文一周后的转、评、赞值进行预测，结果数据表如表8所示，

|  |  |  |
| --- | --- | --- |
| 字段 | 字段说明 | 提取说明 |
| uid | 用户标记 | 抽样&字段加密 |
| mid | 博文标记 | 抽样&字段加密 |
| forward\_count | 博文发表一周后的转发数 |  |
| comment\_count | 博文发表一周后的评论数 |  |
| like\_count | 博文发表一周后的赞数 |  |

表7 赛题5结果数据字段表

## 5.3 评价指标

赛题设置三个评价指标，分别为转发偏差、评论偏差与赞偏差。转发偏差如公式(2)所示，

 (2)

其中，表示预测的转发次数，表示真实的转发次数。

评价偏差如公式(3)所示，

 (3)

其中，表示预测的评论次数，表示真实的评论次数。

赞偏差如公式(4)所示，

 (4)

其中，表示预测的赞次数，表示真实的赞次数。

综合三个偏差，得到一个博文的预测准确率如公式(5)所示，

 (5)

 (6)

其中, 表示预测准确率，如公式(5)所示，表示真实次数，如公式(6)所示，如公式(7)所示，

 (7)

 (8)

## 5.4 赛题网址

<https://tianchi.aliyun.com/competition/entrance/231574/information>

# 心跳信号分类预测 | 长期赛，4288，提升空间大

## 6.1 赛题描述

赛题以医疗数据挖掘为背景，要求选手使用提供的心跳信号传感器数据训练模型并完成不同心跳信号的分类的任务。

## 6.2 数据描述

数据来自某平台心电图数据记录，总数据量超过20万，主要为1列心跳信号序列数据，其中每个样本的信号序列采样频次一致，长度相等。为了保证比赛的公平性，将会从中抽取10万条作为训练集，2万条作为测试集A，2万条作为测试集B，同时会对心跳信号类别（label）信息进行脱敏。字段表如表9所示，

|  |  |
| --- | --- |
| Field | Description |
| id | 为心跳信号分配的唯一标识 |
| heartbeat\_signals | 心跳信号序列 |
| label | 心跳信号类别（0、1、2、3） |

表8 赛题6数据字段表

## 6.3 评价指标

选手需提交4种不同心跳信号预测的概率，选手提交结果与实际心跳类型结果进行对比，求预测的概率与真实值差值的绝对值（越小越好）。

具体计算公式如下：

针对某一个信号，若真实值为，模型预测概率值为,那么该模型的平均指标为

例如，心跳信号为1，会通过编码转成[0,1,0,0][0,1,0,0]，预测不同心跳信号概率为[0.1,0.7,0.1,0.1][0.1,0.7,0.1,0.1]，那么这个预测结果的为

## 6.4 赛题网址

<https://tianchi.aliyun.com/competition/entrance/531883/information>

# 一见钟情指数预测 | 长期赛，无排名可言

## 7.1 赛题描述

本次一见钟情学习赛，受到哥伦布亚商学院教授Ray Fisman和Sheena Iyengar联合发布的文章《伴侣选择中的性别差异(Gender Differences in Mate Selection: Evidence From a Speed Dating Experiment)》启发，决定利用机器学习的数据实验来分析当下男女在相亲交友时，个人的内在外在各类因素对最终相亲结果的影响。

在2002年-2004年期间，Ray Fisman教授和Sheena Iyengar教授在筹备论文时，邀请志愿者参加闪电速配实验(相亲车轮战，每4分钟与一名相亲对象快速沟通，然后再换下一个相亲对象)，提供一些相关的个人信息给相亲对象，并询问相亲对象给出是否愿意在不久的未来再次见面。本次学习赛的分析数据，记录了当时一见钟情相亲实验时，志愿者的相关信息及相亲结果。

## 7.2 数据描述

数据集的内容包括实验志愿者的性别、年龄、人种、专业、地区、收入等特征，以及志愿者对配偶是否来自同一地区、同一信仰等观点的预期。

## 7.3 赛题任务

选手可以针对数据集不同字段间的相互影响进行分析，训练一个机器学习模型，去预测实验人身上一个或多个特性对其相亲成功与否的影响。也就是利用其它特征信息，预测数据集中的“match”字段的结果，1=成功，0=不成功。

## 7.4 赛题网址

[https://tianchi.aliyun.com/competition/entrance/531825/information](https://www.datafountain.cn/competitions/542)

# 贷款违约预测 | 长期赛11178，感觉很有意思

## 8.1 赛题描述

以金融风控中的个人信贷为背景，根据贷款申请人的数据信息预测其是否有违约的可能，以此判断是否通过此项贷款，这是一个典型的分类问题。

## 8.2 数据描述

该数据来自某信贷平台的贷款记录，总数据量超过120w，包含47列变量信息，其中15列为匿名变量。为了保证比赛的公平性，将会从中抽取80万条作为训练集，20万条作为测试集A，20万条作为测试集B，同时会对employmentTitle、purpose、postCode和title等信息进行脱敏。字段表如下：

| Field | Description |
| --- | --- |
| 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 |

表9 赛题8数据描述

## 8.3 评价指标

提交结果为每个测试样本是1的概率，也就是y为1的概率。评价方法为AUC评估模型效果（越大越好）。

## 8.4 赛题网址

<https://tianchi.aliyun.com/competition/entrance/531830/information>

# 挖掘幸福感 | 长期赛，8610，有最优解

## 9.1 赛题背景

在社会科学领域，幸福感的研究占有重要的位置。这个涉及了哲学、心理学、社会学、经济学等多方学科的话题复杂而有趣；同时与大家生活息息相关，每个人对幸福感都有自己的衡量标准。如果能发现影响幸福感的共性，生活中是不是将多一些乐趣；如果能找到影响幸福感的政策因素，便能优化资源配置来提升国民的幸福感。目前社会科学研究注重变量的可解释性和未来政策的落地，主要采用了线性回归和逻辑回归的方法，在收入、健康、职业、社交关系、休闲方式等经济人口因素；以及政府公共服务、宏观经济环境、税负等宏观因素上有了一系列的推测和发现。

赛题尝试了幸福感预测这一经典课题，希望在现有社会科学研究外有其他维度的算法尝试，结合多学科各自优势，挖掘潜在的影响因素，发现更多可解释、可理解的相关关系。

## 9.2 赛题说明

本题使用公开数据的问卷调查结果，选取其中多组变量，包括个体变量（性别、年龄、地域、职业、健康、婚姻与政治面貌等等）、家庭变量（父母、配偶、子女、家庭资本等等）、社会态度（公平、信用、公共服务等等），来预测其对幸福感的评价。

幸福感预测的准确性不是赛题的唯一目的，更希望大家对变量间的关系、变量群的意义有所探索与收获。

## 9.3 数据描述

考虑到变量个数较多，部分变量间关系复杂，数据分为完整版和精简版两类。可从精简版入手熟悉赛题后，使用完整版挖掘更多信息。complete文件为变量完整版数据，abbr文件为变量精简版数据。

index文件中包含每个变量对应的问卷题目，以及变量取值的含义。

survey文件是数据源的原版问卷，作为补充以方便理解问题背景。

数据来源：赛题使用的数据来自中国人民大学中国调查与数据中心主持之《中国综合社会调查（CGSS）》项目。赛题感谢此机构及其人员提供数据协助。中国综合社会调查为多阶分层抽样的截面面访调查。

外部数据：赛题以数据挖掘和分析为出发点，不限制外部数据的使用，比如宏观经济指标、政府再分配政策等公开数据。

## 9.4 评测指标

提交结果为csv文件，其中包含id和happiness的预测值两列。

分数计算公式：

其中n代表测试集样本数， 代表第i个样本的预测值，y\*代表真实值

## 9.5 赛题网址

<https://tianchi.aliyun.com/competition/entrance/231702/information>

# 安全恶意程序检测 | 长期赛，3090，提升空间大

## 10.1 赛题背景

恶意软件是一种被设计用来对目标计算机造成破坏或者占用目标计算机资源的软件，传统的恶意软件包括蠕虫、木马等，这些恶意软件严重侵犯用户合法权益，甚至将为用户及他人带来巨大的经济或其他形式的利益损失。近年来随着虚拟货币进入大众视野，挖矿类的恶意程序也开始大量涌现，黑客通过入侵恶意挖矿程序获取巨额收益。当前恶意软件的检测技术主要有特征码检测、行为检测和启发式检测等，配合使用机器学习可以在一定程度上提高泛化能力，提升恶意样本的识别率。

## 10.2 赛题说明

本题目提供的数据来自文件（windows 可执行程序）经过沙箱程序模拟运行后的API指令序列，全为windows二进制可执行程序，经过脱敏处理。

本题目提供的样本数据均来自于从互联网。其中恶意文件的类型有感染型病毒、木马程序、挖矿程序、DDOS木马、勒索病毒等，数据总计6亿条。

## 10.3 数据描述

训练数据（train.zip）：调用记录近9000万次，文件1万多个（以文件编号汇总），字段描述如下：

|  |  |  |
| --- | --- | --- |
| 字段 | 类型 | 解释 |
| file\_id | bigint | 文件编号 |
| Label | bigint | 文件标签，0-正常/1-勒索病毒/2-挖矿程序/3-DDoS木马/4-蠕虫病毒/5-感染型病毒/6-后门程序/7-木马程序 |
| Api | String | 文件调用的API名称 |
| Tid | bigint | 调用API的线程编号 |
| Index | String | 线程中API调用的顺序编号 |

表10 赛题10数据描述

注1：一个文件调用的api数量有可能很多，对于一个tid中调用超过5000个api的文件，我们进行了截断，按照顺序保留了每个tid前5000个api的记录。

注2：不同线程tid之间没有顺序关系，同一个tid里的index由小到大代表调用的先后顺序关系。

注3：index是单个文件在沙箱执行时的全局顺序，由于沙箱执行时间有精度限制，所以会出现一个index上出现同线程或者不同线程都在执行多次api的情况，可以保证同tid内部的顺序，但不保证连续。

2）测试数据（test.zip）：调用记录近8000万次，文件1万多个。

说明：格式除了没有label字段，其他数据规格与训练数据一致。

## 10.4 评测指标

1.选手的结果文件包含9个字段：file\_id(bigint)、和八个分类的预测概率prob0, prob1, prob2, prob3, prob4, prob5 ,prob6,prob7 (类型double，范围在[0,1]之间，精度保留小数点后5位，prob[removed]=1.0我们会替换为1.0-1e-6)。选手必须保证每一行的|prob0+prob1+prob2+prob3+prob4+prob5+prob6+prob7-1.0|<1e-6，且将列名按如下顺序写入提交结果文件的第一行，作为表头：file\_id,prob0,prob1,prob2,prob3,prob4,prob5,prob6,prob7。

2.分数采用logloss计算公式如下：

M代表分类数，N代表测试集样本数，代表第i个样本是否为类别j(是~1，否~0)，代表选手提交的第i个样本被预测为类别j的概率(prob)，最终公布的logloss保留小数点后6位。

## 10.5 赛题网址

[https://tianchi.aliyun.com/competition/entrance/231694/information](%20https://tianchi.aliyun.com/competition/entrance/231694/information)