实验报告

1.实验简介

Kickstarter作为一家专为具有创意方案的企业筹资的众筹网站平台，致力于支持和激励创新性、创造性、创意性的活动，通过网络平台面对公众募集小额资金，让有创造力的人有可能获得他们所需要的资金，以便使他们的梦想实现。该平台的用户一方是有创新意渴望进行创作和创造的人，另一方则是愿意为他们出资金的人，以此作为连接。

本实验通过抓取数据、回归分析并建立模型，聚焦于来自亚洲且2021年10月前已结束的科技类众筹项目，预测模型的输出为是否成功，即筹集金额是否达到目标。

2.实验目标

本次实验的目标有两个：

（1）从Kickstarter抓取众筹项目数据；

（2）基于抓取的数据，建立对众筹项目成功与否的预测模型

3.实施阶段

3.1数据爬取

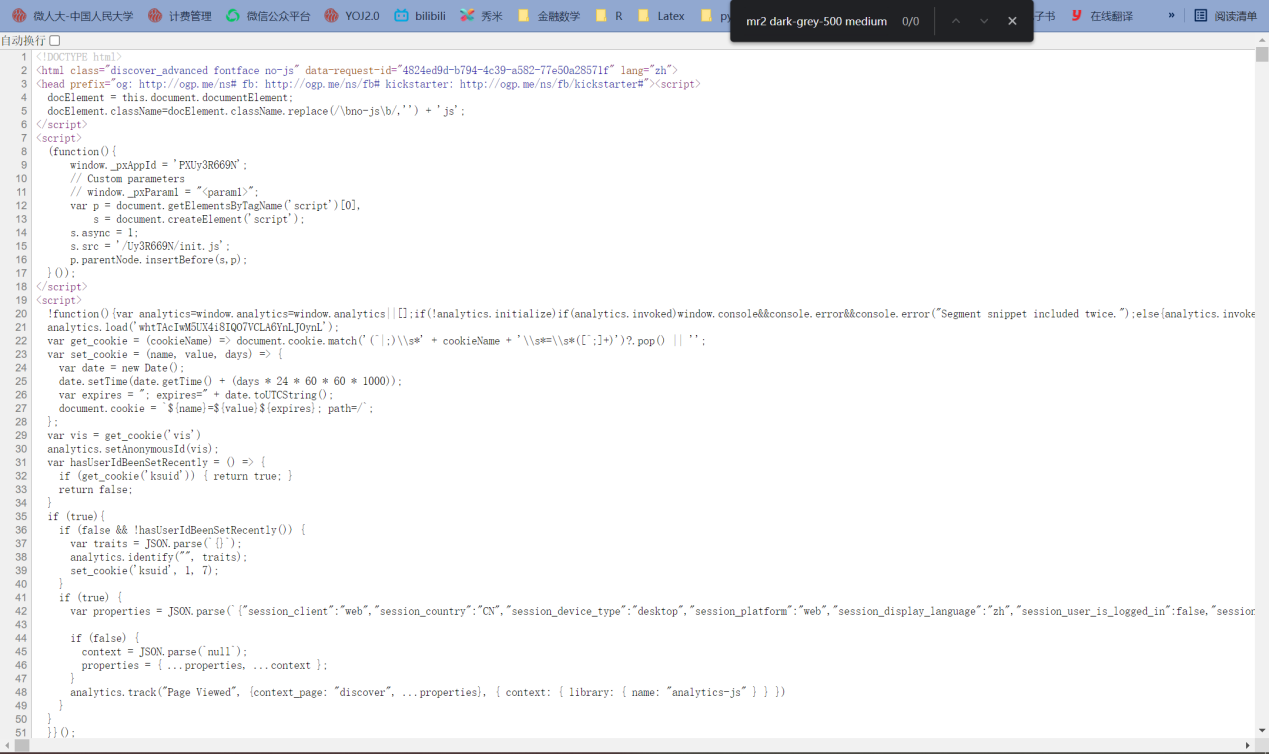
3.1.1网页分析

我们注意到，尽管kickstart是异步加载的网站，但是在其发生异步加载时网址发生了变化。初始网址为：<https://www.kickstarter.com/discover/advanced?category_id=16&woe_id=24865671&sort=magic&seed=2722523>。当点击【加载更多】时，网址变为：<https://www.kickstarter.com/discover/advanced?category_id=16&woe_id=24865671&sort=magic&seed=2722523&page=2>。在尾部多了一个【&page=2】，因此我们就通过更改page的值实现1971（截至爬取时）个提案的网址获取。进一步观察得知，每一个page下共有12个提案，因此page共有1971/12=164.25，将其向上取整即为我们需要循环的次数。

确定好异步加载的的问题后，我们在通过浏览器的【检查】功能查找所需信息时，发现如筹款率的信息的位置如图所示：



当我们进入源代码界面以临近的关键词查找时，却无法找到：



我们怀疑是反扒机制导致有效信息不会通过这种方式展现出来。但我们调整标头信息后依然未果。经师姐提醒后，我们得知，有效信息可能不会以检查中的形式展现出来，但它仍藏在源代码中。故我们尝试采用requests和beautifulsoup方法提取信息：



我们发现，项目名称，地点，筹款率，目标等信息均以类似字典的形式藏在其中。

3.1.2 爬取思路

综上探索过程，我们确定出如下爬取数据的思路：

|  |  |  |
| --- | --- | --- |
| S1 | For循环 | 依次访问165页page，获取该page的HTML文本 |
| S2 | Beautifulsoup | 根据标签和属性，利用find\_all()方法提取存有12个项目信息的HTML文本 |
| S3 | For循环 | 遍历12个项目 |
| S4 | 正则表达式 | 利用正则表达式，提取出项目的json数据 |
| S5 | Json库 | 使用json.loads()方法，将json数据转化为字典类型 |
| S6 | 列表推导式 | 提取出字典的所有值 |
| S7 | Numpy库与pandas库 | 将数据转化为pandas的DataFrame类型，并保存为Excel文件 |

按照上述思路，我们编写了一个Crawl类，以“page=80”为分界，由两人分头爬取80前和80后的页面数据。代码详见附件。

3.2数据清洗

3.2.1变量剔除

抓取下来的原始数据总共有37个变量，显然里面有大量对预测结果无关的变量。所以我们酌情进行剔除。

3.2.2时间

原始数据中的四个时间变量的数据都是十位数字的时间戳类型，并不显示年月日，所以我们使用pandas库的to\_datetime()方法将其转化为datetime类型，并根据deadline变量，只保留了2021年10月前已结束的项目数据。

此外，我们还将deadline与launched\_变量相减，构造出period即筹款周期这一新变量。

3.2.3 location与category

项目展示的定位地址信息保存在location变量里，但是location的数据是json类型，且包含的定位信息复杂多样。我们使用正则表达式，提取出“expanded\_country”的信息，即定位所在的国家全名，而过滤掉其他定位信息。

变量category的数据也是json类型，同上，我们只保留了“name”信息。

此外，location和category作为分类变量，其子类别的数量分布极不均衡，部分子类别的项目数量不足十个甚至只有一个，所以对于项目数量小于10的子类别，我们又进行了剔除，最终两者各保留了12个子类别。

最后，我们对二者进行了编码，并各构造了11个二值虚拟变量，从而进行预测。

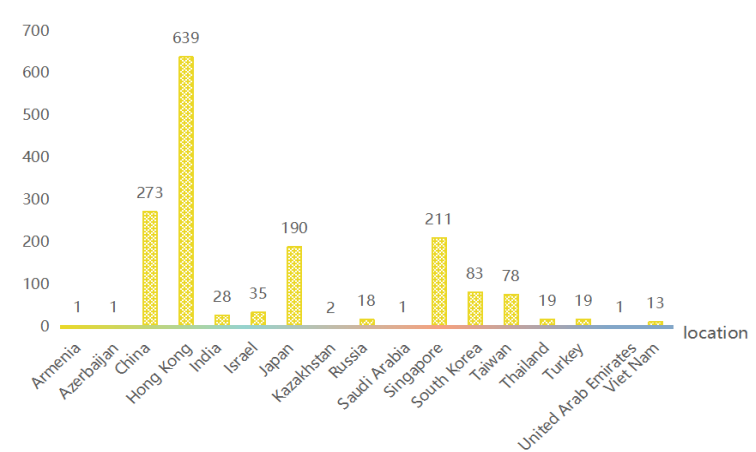
3.2.4构建变量success

由于我们是对是否成功进行预测，所以我们又依据percent\_funded变量，构造了变量success。其中“1”代表募集成功，“0”代表失败。

3.3 可视化

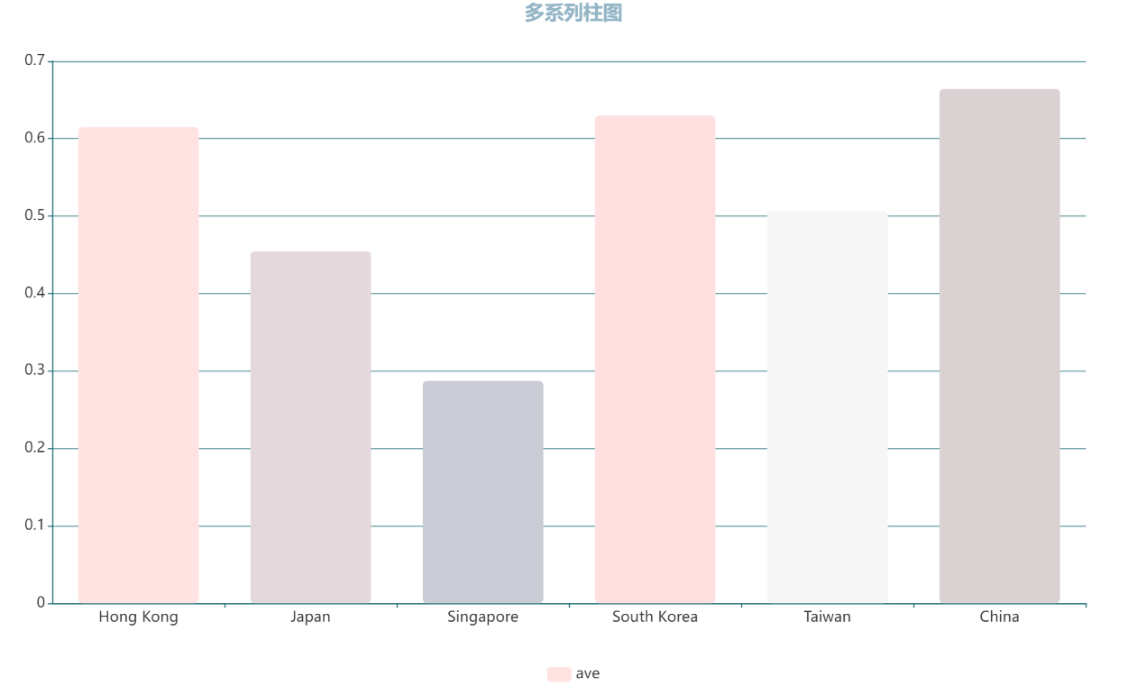
3.3.1 location

从柱状图可以发现，各国家/地区的项目量十分失衡，中国香港以639个独占鳌头，整体上东亚地区的数量最多，但有5个国家的数量为个位数。



各地区筹款项目数

我们从中选取了筹款项目数较多的几个国家（地区），计算出其平均筹款成功率：

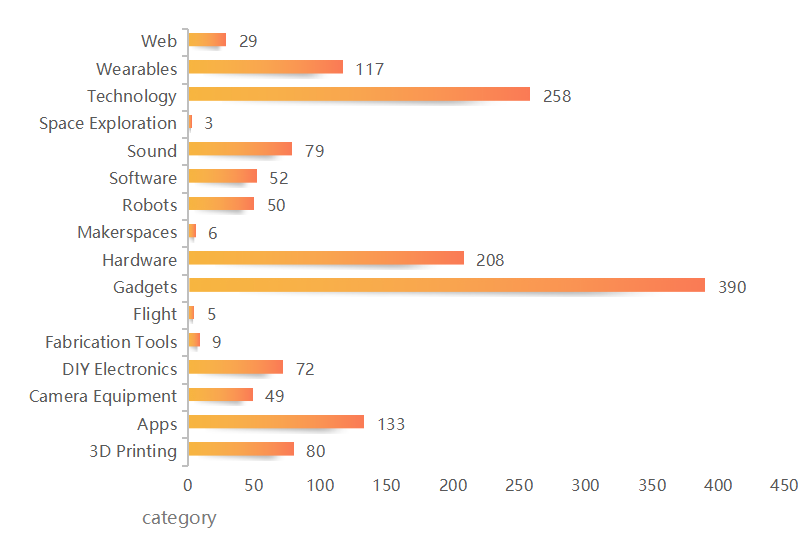


各国家（地区）的平均筹款成功率

我们发现，不同的国家（地区）在筹款成功率上展现出不同的能力。中国，中国香港，韩国表现得较好，而新加坡则不尽人意。故在建立模型时考虑将location因素纳入。

3.3.2 category

从不同种类项目数的条形图可以发现，项目以硬件设备为主，，Software、Web和APPs等软件项目占比居少。



不同种类的筹款项目数

3.4预测

我们使用XGBoost模型来进行项目成功与否的预测。

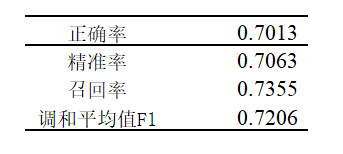
3.4.1输入

进过前期的大清洗，我们最后选择输入模型的变量详见下表。

|  |  |  |
| --- | --- | --- |
| 输入变量 | 类型 | 含义 |
| goal | 连续性变量 | 目标金额 |
| period | 连续性变量 | 筹款周期 |
| staff\_pick | 虚拟变量 | 是否被选为「我们喜爱的专案」 |
| loci (i=0,1,...,10) | 虚拟变量 | 表示经剔除后的12个国家/地区 |
| catei (i=0,1,...,10) | 虚拟变量 | 表示经剔除后的12个类别 |

3.4.2预测效果

按照7：3划分训练集和测试集，得到训练集的正确率为93%，测试集的评价效果如下表所示：



3.4.3其他

关于XGBoost原理以及结果分析详见PPT。