登峰杯

论文类别	☐ 学术作品—自然科学类　☐ 学术作品—人文社科类 ☐ 数学建模竞赛　☑ 数据挖掘竞赛　☐ 艺术创意设计竞赛
论文题目	331100234-教育众筹项目成败分析与预测

清华大学教育研究院
中国高等教育学会学习科学研究分会

教育众筹成败项目分析与预测

摘 要

近年来，互联网众筹在世界范围内流行开来。数十万来自全球各地的老师在教育众筹网站上发出请求书。因此，面对庞大的数据量，分析并预测众筹的成功与否便显得很有意义。

对于问题一，通过 pyhton 自然语言知识处理并提取关键词，利用 panda 模块编程完成了读取文本信息到循环读取并提取到最后输出的过程。结合众筹实际与关键词分析，分析出描述文本与需求文本是大众通常所会关注的。

对于问题二，利用 excel 对原始数据集进行数据选择、数据预处理、数据合并与量化。对各因素进行分类整合之后使用统计方法计算出各类别与成功率得关系，找出了学科、地区、总金额在内的六个影响因素并分析了原因。

对于问题三，根据数据集的特点选择建立分类模型进行成败预测。在 Matlab 中实现了 KNN 算法与 Cart 决策树算法的预测，准确率分别为 73.4%与 83.4%。改进模型后使用组合分类器——Adaboost 算法进行预测，分类器评估准确率为 85.4%。经过比较选择 Adaboost 分类器进行预测。

对于问题四，在第二问与第三问的基础上通过信息熵计算得出学科、历史众筹次数与总金额为影响最大的三个因素。之后详细讨论了这三个因素对于大众决策的影响，最后对众筹平台与老师分别提出合理建议。

关键词： python nltk KNN 算法 Cart 决策树 Adaboost 算法 信息熵

目录

1、挖掘目标

本次数据挖掘目标是利用互联网教育众筹数据，利用 python 自然语言提取目标文本关键词、Adaboost 分类器及信息熵计算，达到以下三个目标：

（1）利用文本分词与词性标注，实现文本的关键词提取，从而找到一种可以使大众快速了解请愿书上的内容的方法。结合众筹方式特点与生活实际研究受欢迎的众筹项目所需各种条件，并讨论大众关注的信息。

（2）根据互联网教育众筹数据，建立分类模型，预测一份请愿书是否会被认可和通过，帮助众筹平台节省审核成本，关注更高质量的需求。

（3）利用信息熵计算分析众筹数据，找到对请愿书是否被大众支持的影响最大的三个因素，具体分析其在影响大众决策时起的作用，给老师发布更好的众筹项目提出建议。

2、分析方法与过程

2.1 数据准备

2.1.1 数据选取

在互联网众筹原始数据集中，有接近二十项的分类数据，总大小接近 600MB。如何从其中选取本次数据挖掘所需要的数据是必须解决的问题。通过登陆众筹网及网络查询研究，我们找出对请愿书是否被认可可能有影响的 6 项除文本以外的因素，分别是：teacher_prefix（教师性别）、school_state（地区）、project_grade_category（受帮助者的年级）、project_subject_categories（项目科目类别）、teacher_number_of_previously_posted_projects（历史众筹申请次数）与 quantity 乘以 price（总众筹金额）。对于其余各项例如教师 id 号码、所需具体物品一类错综

复杂，考虑时个人主观性因素过强，不适于研究大众普遍心理作用，故未列入本次挖掘的范围内

2.1.2 数据清洗

题目所给的 data.csv 的十八万条数据中，出现了许多错位、空白的数据。我们利用 EXCEl 软件对原始数据进行筛选，对其中有问题的数据进行了修正，对无法修正的数据进行了删除。例如在教师。我们去除了约 1 万条左右的不完全数据、噪声数据以及矛盾数据等不适合用来训练和学习的数据，使得数据总数下降至十七万七千条。

2.1.3 数据整合

题目给出了 data.csv 与 resources.csv 两个表格，分别写有同一个众筹项目（id 号码）的不同信息。第二个表格写有所需资源与数量，单价。考虑到众筹时投资方可能更看重众筹金额且所需具体资源各不相同，所以我们决定只读取表二的金额部分并合并到第一个表格中。我们首先利用 excel 对第二个表格中每一个 id 号码中的各项需求所需的金额进行加和，利用工作站在 python 下将上百万条数据按照 id 号码合并为 26 万条数据。这说明表二所含有的项目个数大于表一。所以便调用 excel 的 vlookup 函数在表二中找出表一中所含有的项目并将其合并到表一中，完成合并。

2.1.4 数据构建

为了提高数据挖掘的质量，我们对数据集中错综复杂的各种非文本属性进行整理。由于数据大部分为文字属性，所以我们将文字描述分类并赋值进行量化处理。例如教师性别里男性赋为 1，未婚女性（MS）为 2，MRS 为 3。对于所需金额则按照每 200 美

元为一个区间分别赋值。对于地区我们则是按照美国本土四个时区与海外领地分为五类。在处理时我们将文本部分暂时分出，单列为一个表格单独处理。赋值原则与部分整合结果请看下图（完整表格请见附录 Data.csv）

教师性别	赋值	年级	赋值	科目	赋值	历史众筹次	赋值	众筹金额	赋值	state	赋值
MR	1	pre-K-2	1	applied scie	1	0-4	1	0--200	1	太平洋时区	1
MRS	2	3--5	2	Health&spo	2	5--9	2	200-400	2	山区时区	2
MS	3	6--8	3	History	3	10--14	3	400-600	3	中央时区	3
		9--12	4	literacy	4	15--20	4	600-800	4	东岸时区	4
				Math&Scien	5	20--	5	800-1000	5	海外	5
				Music	6			1000-1200	6		
				Social	7			1200-1400	7		
				warmth	8			1400-1600	8		
								1600-1800	9		
								1800-	10		

表 1

ID	SEX	STATE	GRADE	SUBJECT	NUMBER AP	TOTAL MON	SUCCESS
p141044	1	3	4	2	1	2	1
p162151	2	3	2	2	5	4	1
p235038	2	1	1	5	1	2	0
p059958	3	1	1	5	2	1	1
p093836	1	4	1	4	5	1	1
p167866	3	3	1	7	1	6	1
p075473	3	3	1	1	1	5	0
p062605	2	3	2	4	1	2	1
p236525	3	2	2	2	1	4	1
p097670	3	4	1	1	1	3	1
p231672	2	4	1	5	1	7	0
p143084	2	4	2	4	1	1	0
p214423	3	4	4	5	1	7	1
p165673	1	3	2	5	1	2	1
p044036	2	3	2	2	1	4	1
p116209	3	3	4	3	4	3	1
p151685	2	3	3	5	1	2	1
p046934	2	4	1	4	1	7	1
p168224	2	3	2	5	1	2	1

表 2

至此我们便完成了数据的准备工作。

2.2 问题一的分析方法与过程

从文本中提取关键词的常用工具是 python，文本数据挖掘也是当前的热点方向。自然语言处理也在我们的生活中应用广泛：从电子词典的自动翻译到语法自动纠错，从几乎所有智能手机都有的语音助手到人机交互。基于此，我们充分利用互联网资源

与相关书籍，利用所学改进了寻找到的提取关键词的核心代码。其原理是先利用正则表达式匹配计算机无法理解的文本信息，建立正则词性标注器。利用自然语言拓展包中的布朗语料库进行训练后标注目标文本，正确率理论上可达到 90%。然后利用英文语法知识对目标文本中词频较大的关键名词动词等进行了提取，形成关键词列表。为了可以处理大量的文本，我们调用 panda 模块进行编程使之可以循环处理表格中的文本数据并输出。

2.2.1 流程图

2.2.2 循环提取：

我们编写的 panda 循环提取代码如下：

```
import pandas as pd
import numpy as np
a = pd.read_csv('D:\data.csv')
b = a['project_essay_1']
size = np.size(b)
result_all = pd.DataFrame()
for i in range(size):
    sentence = b[i]
    np_extractor = NPExtractor(sentence)
    result = np_extractor.extract()
    data = pd.DataFrame(result)
    result_all = pd.concat([result_all, data], axis=1)
result_all.to_csv('D:/test.csv')
```

（引用的关键词提取程序请见附录 程序.py。）

2.2.3 提取关键词

考虑到文本数据非常庞大，我们决定先分出四篇 essay 部分里的 essay1 部分进行关键词提取测试。提取结果比较理想，例如下表中的第一个项目就提取出了如‘low income’，‘hard work’等关键词，能比较好的反映文本信息。再对其他 esaay 进行提取后，综合结合其它非文本信息即可使人快速了解该众筹项目。

id	0	1	2	3	4	5	6	7	8
p036502	kindergarter	low-income	English	kindergarter	school settir	new things	sight words	letter sound	hard work
p039565	elementary s	rich school	diverse popu	Pre-K	Title	school popu	high concen	English Lear	foster group
p233823	Hello	\r\nMy nam	Mrs. Brother	5th grade	Ascent	wonderful cl	wonderful te	\r\nMy stud	gifted kids
p185307	inner city sc	PE	Physical	physical acti	good progra	African	Hispanic	students ran	kindergarter
p013780	physical acti	elective clas	own food	healthy meals					
p063374	happy teach	Respectful	On-Task	Responsible	Safe	school 's cor	graders ama	school year	months.\r\n
p103285	Kindergarter	high expecta	Kindergarter	attention sp	excess energ	Minecraft	Barbies	huge imagir	diverse soci

表 3

（详见关键词.csv）

2.3 问题二的分析过程

除了文本信息之外，还有许多因素会影响请求书的被认可度，例如众筹的金额，众筹的学科方面等等。基于数据准备工作时已经完成的数据各因素分类表，我们利用 Excel 软件对每一项因素对成功率的影响进行了分析

2.3.1 数据统计

利用 excel 软件的筛选和绘图功能对每一个因素的每一类进行成功率计算并展示，可以直观的看出成功率随因素变化而发生的改变。据此找出对众筹影响较大的因素。例如对众筹金额一项的分析中可以明显看出成功率随金额上升而先下降后上升。

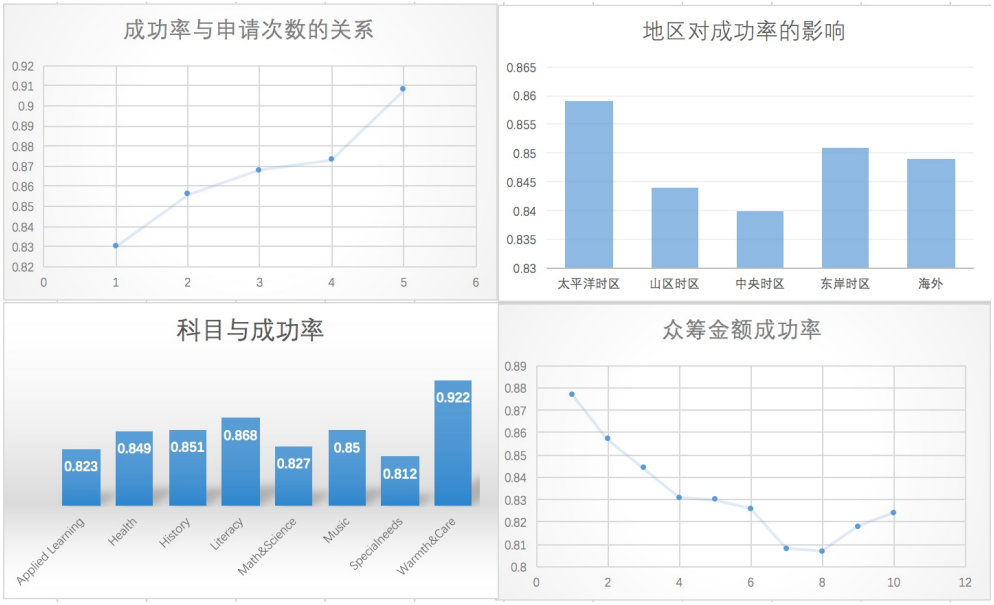


图 1

使用这种方法验证了数据准备时提取的六项因素对成功率都有一定影响。所以保留这六项因素作为问题三模型的数据属性。

2.4 问题三的分析方法与过程：

2.4.1 问题三流程图：

2.4.2. 数据预处理

2.4.2.1 数据抽样

为了既加快运算速度，又要保证训练精度，我们从整理好的十七万条数据中随机抽取了 10%即 17750 条数据用于数学建模。在抽取数据时我们采用了添加随机数的抽样方法，即（1）在每一项的后面都产生一个 6 位随机数（2）根据随机数的大小对原数据进行排序（3）取排列后的数据集的前 17750 项作为新的数据集。

2.4.2.2 划分训练集

为了有效利用数据中的规律并进行训练，我们将数据集划分为两部分，一部分作为训练集，一部分作为测试集。在这里，为了保证训练样本足够多，训练集占原数据集 90%，测试集占 10%

2.4.3 构建分类模型

数据集的特点是离散的类型数据，结果为成功和不成功两类。根据数据集的特点，我们决定构建分类模型进行预测。

2.4.3.1 KNN 临近算法

我们首先建立了 KNN 分类模型，KNN 模型操作相对简单，其简介如下：

K 最近邻(k-Nearest Neighbor，KNN)分类算法，是一个理论上比较成熟的方法，也是最简单的机器学习算法之一。该方法的思路是：如果一个样本在特征空间中的 k 个最相似(即特征空间中最邻近)的样本中的大多数属于某一个类别，则该样本也属于这个类别。

KNN 的工作原理如下：

1. 假设有一个带有标签的样本数据集（训练样本集），其中包含每条数据与所属分类的对应关系。
2. 输入没有标签的新数据后,将新数据的每个特征与样本集中数据对应的特征进行比较。
    1. 计算新数据与样本数据集中每条数据的距离。
    2. 对求得的所有距离进行排序（从小到大，越小表示越相似）。
    3. 取前 k （k 一般小于等于 20 ）个样本数据对应的分类标签。
3. 求 k 个数据中出现次数最多的分类标签作为新数据的分类

在 Matlab 中使用 knnclassify 函数对训练集进行训练，得到 KNN 分类器，核心代码如下：

```
%knn聚类
 z1=knnclassify(x1,x,y);
 zc1=1-sum(abs(y1-z1))/1000;%计算误差
 %73.4
```

然后对测试集进行预测，准确率可以达到 73.4%

2.4.3.2 Cart 决策树算法

基于 KNN 分类器分类准确率不是很理想，所以决定采用经典的分类模型：决策树决策树是主要针对“以离散数据作为类型进行分类”的学习方法，它的构造不需要任何领域知识或参数设置，因此适于探究式知识的发现，并且可以处理高维数据，所以非常适合本问的模型构建。Cart 决策树通过建立二叉决策树有效提升了决策树在节点比较多的情况下的运算速度，其简介如下：

CART 分类回归树是一种典型的二叉决策树，可以做分类或者回归。如果待预测结果是离散型数据，则 CART 生成分类决策树；如果待预测结果是连续型数据，则 CART 生成回归决策树。数据对象的属性特征为离散型或连续型，并不是区别分类树与回归树的标准，作为分类决策树时，待预测样本落至某一叶子节点，则输出该叶子节点中所有样本所属类别最多的那一类（即叶子节点中的样本可能不是属于同一个类别，则多数为主）；作为回归决策树时，待预测样本落至某一叶子节点，则输出该叶子节点中所有样本的均值。

以下是算法描述：其中 T 代表当前样本集，当前候选属性集用 T_attributelist 表示。

（1）创建根节点 N

（2）为 N 分配类别

（3）if T 都属于同一类别 or T 中只剩下 一个样本则返回 N 为叶节点，否则为其分配属性

（4）for each T_attributelist 中属性执行该属性上的一个划分，计算此划分的 GINI 系数

（5）N 的测试属性 test_attribute=T_attributelist 中最小 GINI 系数的属性

（6）划分 T 得到 T1 T2 子集

（7）对于 T1 重复（1）-（6）

（8）对于 T2 重复（1）-（6）

基于此，利用 Matlab 的 fitctree 函数进行模型构建，在互联网上学习 cart 决策树构建方法并应用，得出了一个有 750 多个节点的决策树。

所生成的决策树如下图所示。

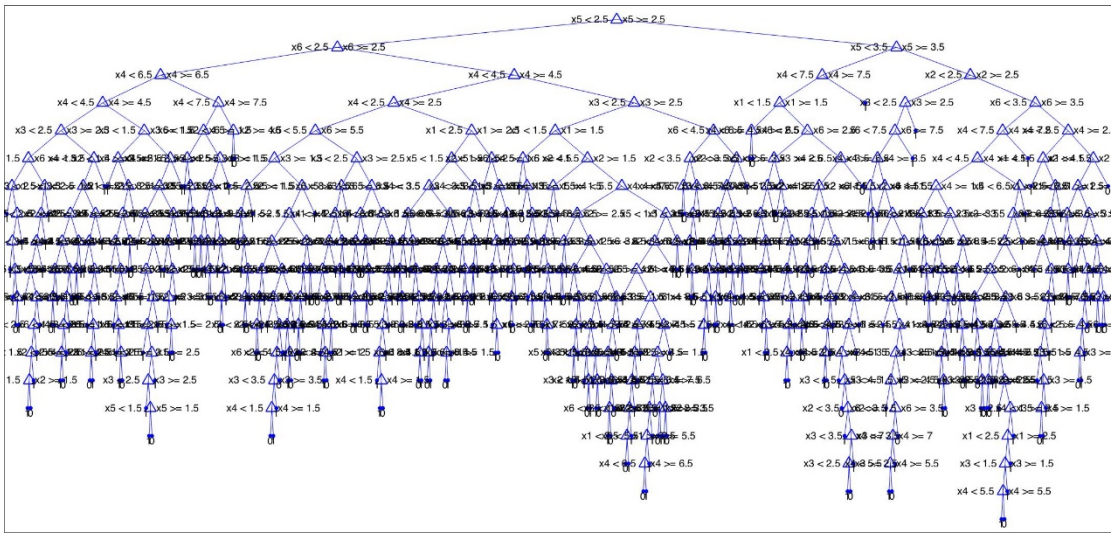


图 2

核心代码如下：

```
%构建CART算法分类树
tree=fitctree(X_train,Y_train);
view(tree,'Mode','graph');%生成树图
rules_num=(tree.IsBranchNode==0);
rules_num=sum(rules_num);%求取规则数量
Cart_result=predict(tree,X_test);%使用测试样本进行验证
%统计准确率
Cart_rate=1-sum(abs(Y_test-Cart_result))/m
disp(['规则数: ' num2str(rules_num)]);
disp(['测试样本识别准确率: ' num2str(Cart_rate)]);
```

对测试集进行保持方法评估分类器准确率可得：准确率为 83.4%，比 KNN 算法准确率明显提升。

2.4.3.3 Adaboost 分类器

题目数据集数据多，维数高，非常复杂。为了进一步提高准确率，我们又尝试了与之前两种模型都不同的组合分类器——Adaboost 分类器，希望可以通过多个弱分类器的组合产生有更高精确度的强分类器。Adaboost 分类器模型简介如下：

Adaboost 是一种迭代算法，其核心思想是针对同一个训练集训练不同的分类器(弱分类器)，然后把这些弱分类器集合起来，构成一个更强的最终分类器(强分类器)。其算法本身是通过改变数据分布来实现的，它根据每次训练集之中每个样本的分类是否正确，以及上次的总体分类的准确率，来确定每个样本的权值。将修改过权值的新数据集送给下层分类器进行训练，最后将每次训练得到的分类器最后融合起来，作为最后的决策分类器。使用 Adaboost 分类器可以排除一些不必要的训练数据特征，并将关键放在关键的训练数据上面。

AdaBoost 算法的伪代码如下：

函数 AdaBoost（D，T）

输入：样本数据集 D，学习提升轮数 T

输出：集成分类器 H（x）

（1）初始化 N 个样本的权重

$$W_1(x) = 1\backslash N(i = 1,2,\ldots,\ N)$$

（2）for t=1 to T do

（3）根据权重W_1的分布，通过对D进行有放回抽样产生训练集D_t

（4）在D_t上训练产生一个弱学习器（基学习器）h_t

（5）用h_t对原训练集D中的所有样本进行分类，并度量h_t的误差

$$e_t = \frac{1}{N}\left[\sum_{i=1}^{N} W_t(i) I(h_t(x_i) \neq y_i)\right]$$

（如果$h_t(x_i) \neq y_i$为真，则$I(h_t(x_i) \neq y_i)$=1，否则为 0）

（6）if$e_t > 0.5$then

（7）重新将权重初始化为$1|N$,转步骤（3 重试）

（8）end if

（9）决定h_t得到权重

$$\alpha_t = \frac{1}{N}\ln\left(\frac{1-e_t}{e_t}\right)$$

（10） 更新权重分布

$$W_t(i) = \frac{W_t(i)}{Z_i} \times \begin{cases} \exp(-\alpha_t) \text{ if } h_t(x_i) = y_i \\ \exp(\alpha_t) \text{ if } h_t(x_i) \neq y_i \end{cases} = \frac{W_i \exp(-\alpha_t y_i(x_i))}{Z_i}$$

其中：Z_i是一个正规因子，用来确保$\sum_t W_{t+1} = 1$。

（11） end for

（12） $H_{(x)} = \text{sign}(\sum_{t=1}^{T} \alpha_t h_t(x))$

其中：sign()为符号函数，$\text{sign}(\sum_{t=1}^{T} \alpha_t h_t(x))$ $\begin{cases} 1\ if\ \sum_{t=1}^{T} \alpha_t h_t(x) \geq 0 \\ -1\ if\ \sum_{t=1}^{T} \alpha_t h_t(x) < 0 \end{cases}$

据此便可以编程计算。

从 Matlab 官网上下载了 Adaboost 函数，学习其内容及所给‘example’后，利用该函数进行了预测，预测准确率为 85.4%，核心代码如下：

```
%adaboost
  n=200
[classestimate,model]=adaboost('train',x,y,n);%训练
 zbc=1-sum(abs(y-classestimate))/1000;
  z3=adaboost('apply',x1,model);%预测
zbe=1-sum(abs(y1-z3))/1000;%误差
```

2.4.4 模型比较与选择

K-邻近算法准确率：73.4%

CART 决策树算法准确率：83.4%

Adaboost 组合分类器算法：85.4%

比较三种算法的预测准确率，Adaboost 组合分类器算法准确率最高，故本文采用 Adaboost 组合分类器作为预测众筹成功与否的工具。

2.5 问题四的分析方法与过程

此问题我们可用求信息熵的方法解决。“信息熵”的概念由信息论的鼻祖香农提出，他借用了热力学中“熵”的概念，用数学的方法量化了信息的混乱程度及信源的不确定度。公式为：

$$\mathbf{E} = -\sum_{i=1}^{n} p_i \log_2 p_i$$

所以我们先把系统的初始熵算出，即无限制条件时只根据成功与否算出的信息熵。根据公式可得初始熵 E(U)为 0.61264。当加入了影响因素，或者说增加了一些辅助判断的信息，这时的信息熵我们称为后验熵 E(U,Xi)，Xi 为某个限制变量。用初始熵减去后验熵，得到信息增益 Gain(U,Xi)，它的大小反映的是消除不确定性的结果，也就是限制的强度。若 Xi 变量中有 n 种情况，其中每种情况分别占所有情况的 n1,n2…则我们用加权平均的算法，算出每种情况的信息熵，再加上占所有情况的权重，可算出每种变量的后验熵，从而得到信息增益。

信息增益

E(U, X1)	0.61147	Gains(U, X1)	0.00117
E(U, X2)	0.61185	Gains(U, X2)	0.00079
E(U, X3)	0.61196	Gains(U, X3)	0.00068
E(U, X4)	0.60376	Gains(U, X4)	0.00888
E(U, X5)	0.60821	Gains(U, X5)	0.00442
E(U, X6)	0.60972	Gains(U, X6)	0.00291

表 4

从信息增益大小可得，科目，申请次数最重要，其次是钱数和性别，最后是年级和州。

3、结果分析

3.1 问题一结果分析

使用改进后的提取关键词软件对十八万条原数据的 essay 部分进行了关键词提取，并将结果导入至关键词.csv 中。部分关键词如下图所示：

p116102	low-income	meaningful	special neec	\r\n\r\nMy s	positive atte	free lunch	socioeconor	life experien
p070029	high poverty	different cul	African	Latinos	Bangalees	formal educ	Providing	strong found
p107356	bright smile	great start	school year	students gro	bright.\r\nA	diverse learr	BEST	
p031939	school distri	Idaho	strong agric	corn field	front doors	agricultural	large popula	low-income
p044085	community	different lan	low income	free breakfa	school prog	offering pro	food pantry	job assistanc

表 5

通过对众筹网站的模拟参与，我们认为对项目有描述性的文本和需求文本信息会被大众关注。例如上图关键词中 low-income、high poverty 和 African 等就非常有感染力，上图如 free lunch、free breakfast 这些需求文本关键词则可以很清楚的反映需求，更易被大众关注。原始数据集中的 project_resource_summary 一项就是需求文本，为了使大众更快速的了解众筹项目，我们建议将其与 essay 文本提取关键词后搭配上重要的非文本信息如众筹金额、类别等制作成项目简介，再配以图片、动画等

成为项目名片展示在主页上，社会人士可以点击自己感兴趣的项目名片进入下一级查看更具体的信息。

3.2 问题二结果分析

利用 excel 的筛选统计绘图功能对各非文本因素进行了研究，挑选了六个影响因素作为研究属性

3.2.1 众筹金额因素

众筹金额一项的关系图如下（横坐标为赋值后的金额类别）

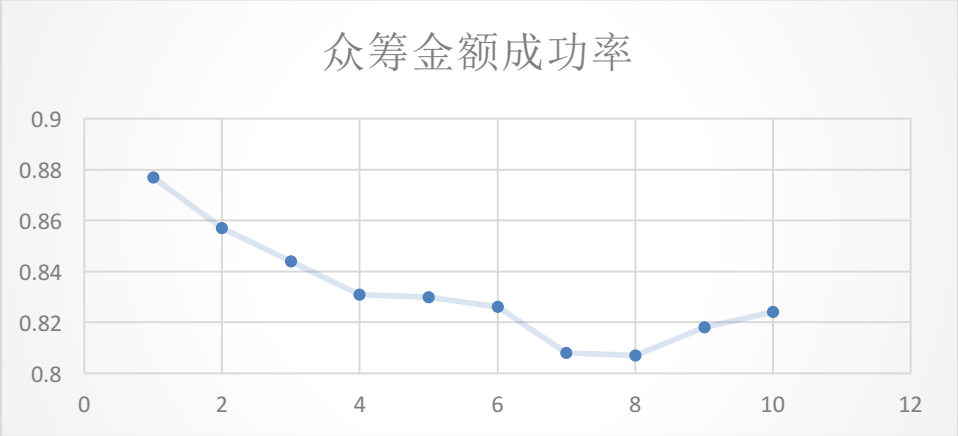


图 3

可以看出这两项对于成功率的影响是很大的。其中众筹金额一项对于其成功率的影响是非常明显的。成功率随着众筹金额的增加先减小后增加，这种现象出乎我们的意料。在结合生活实际与上网调查后，经过分析我们认为这可能是由于在金额较少时，众筹难度较低，成功率高；金额较大时，捐助者认为此项目确实需要众筹帮助，难以自行解决，反而有可能众筹成功。但金额在第六到第八类即 1000-1600 美元时成功率较低，所以建议发起人在考虑金额大小时避开这个范围。

3.2.2 历史众筹次数因素

根据历史众筹次数统计并绘制的各类别平均成功率如下图所示：

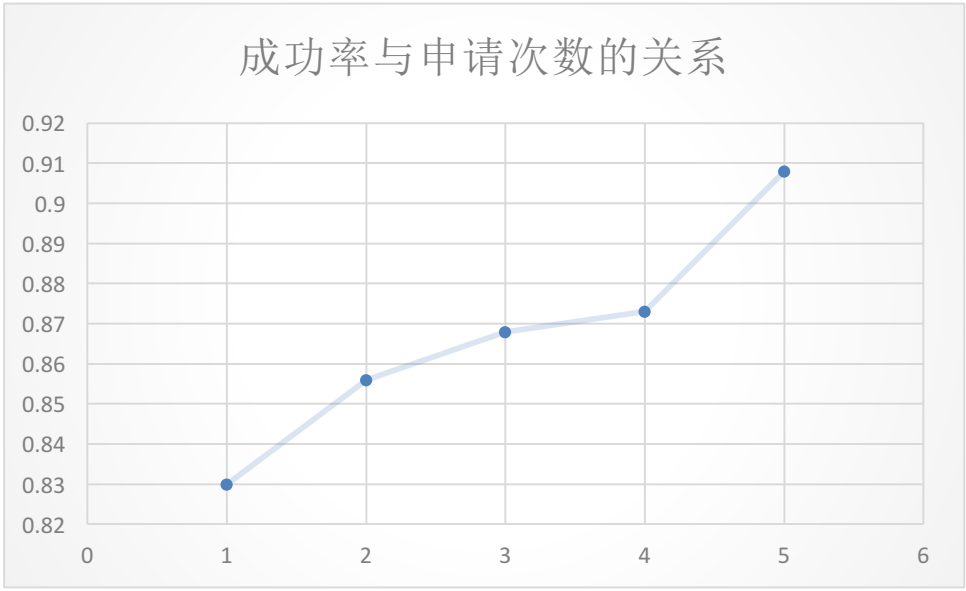
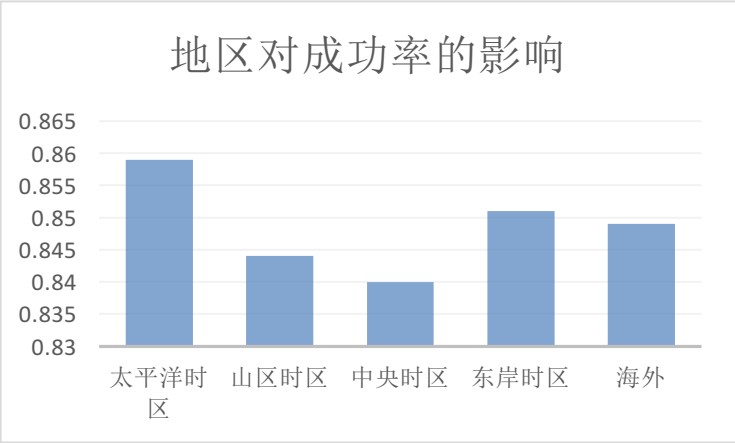


图 4

分析成功与历史众筹次数之间的关系如上图所示，得出历史众筹次数越高，成功率越高。众筹发起人众筹次数越多，经验越丰富，信誉越高，更易被投资者认可。建议发起人可以寻找有经验的老师帮助进行众筹，提高成功率。

3.2.3 地区因素

根据数据处理时按照时区划分的五个地区，可以直观的看出地区对成功率的影响作用



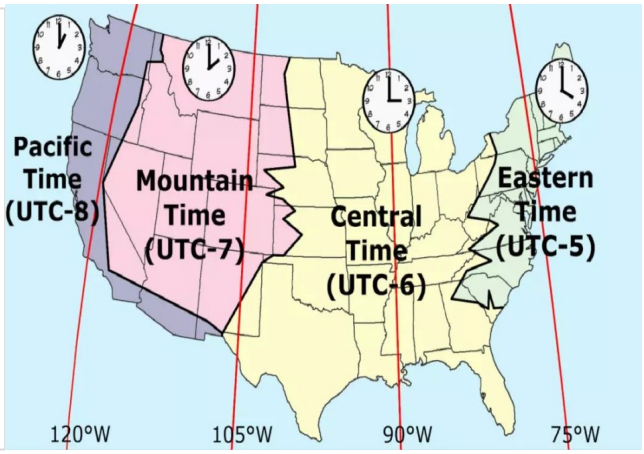
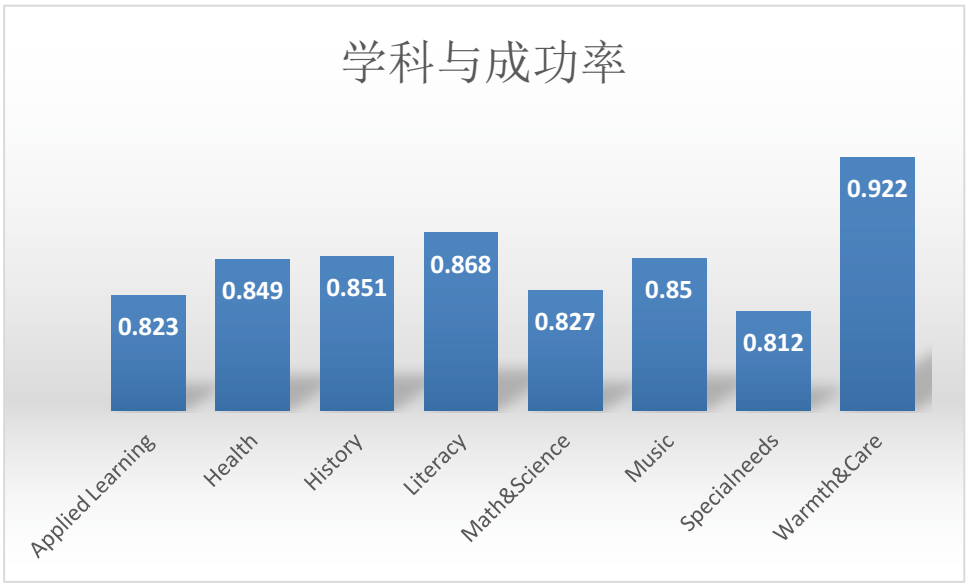


图 5

在分析地区影响因素之前，我们的预期结果是教育资源相对薄弱的美国中部成功率应该更高，就像在中国为山区孩子众筹教育费用一样。但挖掘结果又一次与预期不同，中南部地区成功率反而更低，成功率高的是经济发达的美国沿海地区（太平洋时区与东岸时区）。经过上网查阅美国地区信息了解到，中南部地区以乡村生活闻名，属于传统行业聚集区、农业区；沿海地区经济发达，信息技术产业、高新技术产业发达。研究后认为：中南部地区由于传统行业聚集，教育众筹回报率较低，互联网众筹概念可能接受度不高。而投资者可能更愿意投资、支持本地区教育发展，导致沿海发达地区大量慷慨的投资者参与当地的教育众筹使之成功率较高，中南部地区反而较低。在调查中我们发现中国众筹也存在这种现象，在众筹网上可以看到像北京上海周边的众筹成功时间快，而西北地区众筹项目进展缓慢。这应该是因为经济发达地区众筹后回报率高，失败风险低，投资更有保障，所以成功率高。

3.2.4 众筹学科对成功率的影响



分析影响结果统计得到上图，成功率明显偏高的是 Warmth&Care 类别，说明关于以温暖与关怀为目的众筹项目更易成功，例如贫困区学校提供免费午餐（Free Lunch）项目。数学科学与应用学习成功率的相对低一些。建议众筹发起人将类别向人文关怀靠拢，可以提升成功率。

3.3 问题三结果分析

使用 KNN 邻近算法进行预测效果不是很理想。为了更好的预测的项目成功与否，尝试使用 Cart 决策树模型进行预测，大幅提升预测准确率，最终使用组合分类器 Adaboost 对一万六千条数据进行训练，训练后的分类器预测准确率为 85.4%，较为理想。该分类器具有简单性和稳定性的的特点，便于运用，适合众筹平台使用，能基本满足该平台的需求。该方法通过预测请求书的支持度来节省平台的审核成本，将更多的众筹资源转移到真正需要的地方上去，造福贫困区教育困难的人民。

3.4 问题四结果分析

通过信息熵计算我们得出科目（subject）、历史申请次数和众筹金额是三个最重要的因素。其中，subject 通过给众筹项目分类来影响大众支持。比如在问题二中我们得出 warmth&care 即人文关怀的成功率最高高达 91%，说明大众更愿意帮助关爱儿童的项目，从中获得幸福感。

历史申请次数越高，成功率越高，这是与发起人的众筹经验有关的。发起人历史申请次数越多，众筹经验越多，越了解大众会关注什么，对什么更感兴趣，从而改进请愿书，提高成功率。

众筹总金额对大众的影响较为复杂，随着金额升高而现下降后升高。在金额较低的时候，众筹难度较低，支持者付出成本较低，更容易被支持，例如众筹 10 元的项目的成功率肯定不比众筹 100000 元的项目低。但随着金额升高(>1600 美元)，较大的金额会使大众产生同情，觉得贫困地区的老师确实无法自己解决这项教育问题，确实需要社会帮助，所以支持度反而上升。这个变化的临界点在 1600 美元，此时成功率较低。

众筹发布者可以通过改进这三个方面来增加支持度，获得所需的教育资源。

4、结论

对互联网教育众筹的成败分析，了解大众对于教育众筹的关注点、支持度，对于众筹平台、发起人和社会来说都有重要意义。正所谓教育是民族振兴、社会进步的基石，通过众筹分析预测，可以使教育资源向需要它的地方倾斜，改善贫困地区的教育水平。本文采用 pyhton 自然语言进行请愿书关键词提取，利用信息熵计算寻找关键因素，建立 Adaboost 分类器模型预测众筹成败来分析教育众筹。

由分析结果可以，众筹影响因素有文本因素与非文本因素。通过对文本信息进行

提取关键词并分析，得出描述文本与需求文本通常会被大众所关注，发起人和平台可以通过制作含有关键信息的项目名片来使大众快速了解项目信息，更好的作出决策。

经过信息熵计算得出非文本因素中教育类别、教师历史申请次数与众筹总金额是最关键的三个因素，它们通过不同的方式影响着社会人士的决策。发起人可以通过更改教育项目类别、咨询有较多申请经验的老师与调整众筹金额的方式增加项目被支持度，获得所需的教育资源。

利用历史众筹数据建立 Adaboost 分类器实现对众筹成败的预测。众筹平台可以利用该模型对情愿书进行审核，降低其人力成本，并更好的关注高质量的众筹项目，实现教育资源合理分配。

5、参考文献

1. http://blog.chinaunix.net/xmlrpc.php?id=4216038&r=blog/article&uid=29235952 （adaboost）
2. https://blog.csdn.net/zhihua_oba/article/details/72230427（cart 回归树）
3. 决策树（https://baike.baidu.com/item/决策树/10377049）
4. Cart 树 https://blog.csdn.net/u010356524/article/details/79848624
5. KNN 算法 https://baike.baidu.com/item/k 近邻算法/9512781?fr=aladdin
6. KNN 算法 https://blog.csdn.net/u010859707/article/details/70666596
7. 信息熵 https://blog.csdn.net/chenjunji123456/article/details/52189312
8 许国根，贾瑛.实战大数据 MATLAB 数据挖掘详解与实践 北京：清华大学出版社，2017
9 Steven Bird，Ewan Klein & Edward Loper，Natural Language Processing with Python 国外：O'REILLY

官方网站	www.dengfengbei.com
	Dengfengbeijingsai
微信公众号	
官方 QQ 群	（1）“登峰杯”学术作品学生 QQ 群 571526693 （2）“登峰杯”数学建模学生 QQ 群 571535826 （3）“登峰杯”机器人学生 QQ 群 571540979 （4）“登峰杯”结构设计学生 QQ 群 592858677 （5）“登峰杯”数据挖掘学生 QQ 群

	144821810 （6）“登峰杯”艺术创意设计学生 QQ 群 318850726
官方邮箱	dengfengbei@126.com
联系电话	010-52909593，18310079788 （工作日 9:00~12:00，13:00~17:00）