众筹项目成功率影响因素的实证分析及成功率预测——以 Makuake 为例

李峥昊

西安交通大学　大数据 001

日期：2023 年 2 月 1 日

摘　要

本文从 Makuake 众筹网站中获取了全部 29658 条众筹数据，构建了 Makuake 众筹数据集，并基于 Makuake 数据集分析了众筹项目成功率的影响因素。用机器学习和深度学习中的多种方法构建了众筹项目成功率预测模型，并训练了一个文本分类模型以帮助用户自动填写项目类别。实验结果表明，目标金额对于众筹成功率有着较强的负面影响，众筹选项中的金额设置对于成功率有多种影响。摘要长度、目标金额、活动数量等在预测过程中有着重要影响。文本特征对于准确率的提升较为有限。

关键词： 众筹，成功率，预测模型，文本分类，Makuake

1　介绍

众筹（Crowdfunding）是近年来在网络上流行的一种融资方式，它为那些希望获得资金以创办或发展个人或团体项目的人提供了一种新的选择。众筹项目通常通过众筹平台发布在互联网上，以吸引潜在的支持者对项目进行资助。支持者通常会获得项目的一些奖励，这些奖励的数量和种类取决于支持者所提供的金额。

众筹模式有很多种，包括奖励型众筹、股权型众筹、债权型众筹和捐赠型众筹。奖励型众筹是最常见的模式。奖励型众筹为投资者提供非货币奖励，如资助项目的折扣[1]，资助项目的预购[2]，或其它象征赞赏的奖励[3] (Busse and Gregus, 2020)。奖励型众筹一般来说有两种模式，“All or Nothing”和“All in”（Makuake），“All or Nothing”和“Keep it All”（Kickstarter）。

众筹在网络上的流行主要得益于互联网的普及和社交媒体的发展。这些平台为项目发起者提供了一种方便的方式来宣传自己的项目，同时也为支持者提供了一种方便的方式来了解并支持感兴趣的项目。

然而，众筹并不是没有风险的。一些项目发起者可能会许诺不能兑现的奖励，或者甚至是故意欺骗支持者。为了减少风险，支持者可以选择只支持那些具有较强信誉的众筹平台。此外，支持者还应该加强对项目的了解，确保自己对项目有充分的了解，并对自己的支持进行适当的风险评估。

众筹项目的成功率受到许多因素的影响，总体上可以分为三类：众筹项目本身的特点，网络以及可理解性。其中众筹项目本身特点包括众筹目标金额，最低投资额，项目持续时间和提供的财务信息等 (Lukkarinen et al., 2016)。一般来讲，对于奖励型众筹，众筹目标金额对于成功率有负面影响 (Cumming, Leboeuf, and Schwienbacher, 2020; Zheng et al., 2014)。最低投资金额对成功率没有显著影响 (Agrawal, Catalini, and Goldfarb, 2014)。项目持续时间在欧美对于项目的成功率有负面影响，持续时间较长给了投资者较长的思考时间，以至于投资者们有可能忘记这件事 (Härkönen, 2014; Mollick, 2014)，但是在中国，持续时间对于

[1]资助项目的折扣：可以以“早鸟价”或“尝鲜价”购买商品

[2]资助项目的预购：如手办、周边等商品的发售往往预购-制作-售卖的形式，除了预购，只有少量商品公开售卖

[3]象征赞赏的奖励：非卖品，如「在这世界的角落」中的奖励的明信片；权利，如可以参加“生产支持者会议”

项目成功率有正面影响 (Zheng et al., 2014)。在奖励型众筹中，社交网络的规模与项目的成功率有着显著的正相关关系 (Etter, Grossglauser, and Thiran, 2013)。在可理解性方面，产品类项目相较于服务类项目有更高的成功率，因为消费者们更愿意看到有形的结果 (Belleflamme, Lambert, and Schwienbacher, 2013)。项目更容易被投资者理解对于众筹成功率有着正面影响 (Härkönen, 2014)。

2 数据选取说明

2.1 Makuake 简介

图 1: Makuake 众筹平台

Makuake是日本的一家众筹平台。于 2013 年 8 月作为 *CyberAgent*（日本最大的网络广告代理商）的新事业部成立。自 2017 年 glafit 电助力自行车以 1 亿日元以上的筹集金额破了日本国内众筹记录之后，逐渐成为为日本最大的众筹平台。平台众筹项目囊括了日常用品，时尚单品，餐饮店，日本酒，电影等各类行业的各种产品。代表作有动画「在这世界的角落」，日本酒「雪どけ酒」，耳机「VIE FIT」，电助力自行车「glafit」等等（如图2所示）。

2.2 网页信息说明

Makuake 的项目界面主要分为 3 个页面，主界面，活动界面以及应援界面。

主界面的页首（如图3所示）包含该项目的一些基本信息，如标题、筹集金额、目标金额，facebook 点赞数量等等。

主页面的左侧部分包含介绍的摘要和主体部分，右侧包含项目的类型和可以选择的若干种支援方式。项目类型分为 All in 和 All or nothing。All in 表示不论筹集了多少钱，项目发起者都会收到这笔钱，并且要履行对应的服务，All or nothing 如果没有达到目标金额，则筹集金额全数返还。若干种支援方式，即项目发起人往往会在这里设置若干种不同价位不同组合的选项——以电影「在这世界的角落」为例，如果支付 2160 日元，则可以收到不定期的制作进度、动画草图的邮件，以及一张明信片；如果支付 5400 日元，则在前一档的基础上，可以参加导演举行的“生产支持者会议”。

活动页面（如图4所示），包含项目的进度信息。

应援界面（如图5所示）是支持者对于该项目的应援或是评论，只有参与了众筹的用户可以评论，可以多次评论。

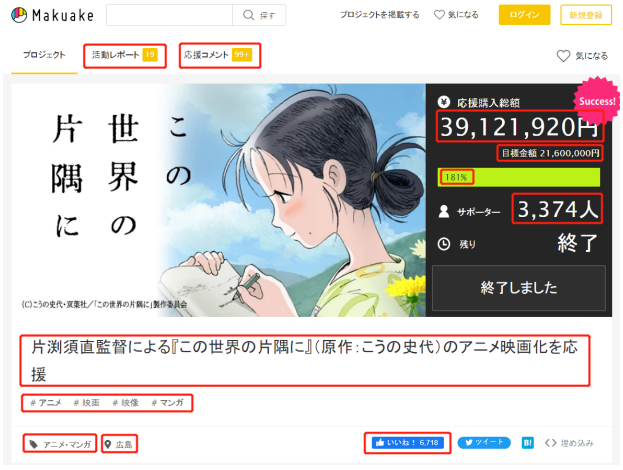
(a) 在这世界的角落

(b) 雪どけ酒

(c) VIE FIT

(d) glafit

图 2: Makuake 代表项目



(a) 页首

(b) 介绍

图 3: 主界面

图 4: 活动页面

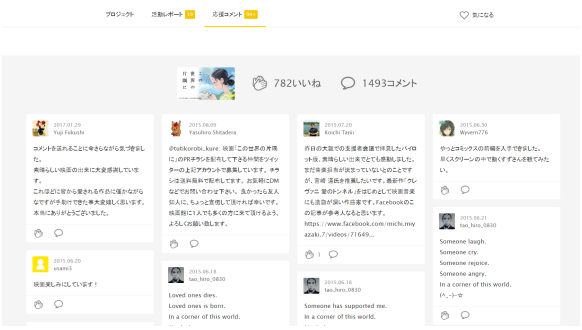


图 5: 应援页面

2.3 数据说明

本次实验数据从 5 个 url【项目主页，investment（项目主页的补充信息），facebook 点赞数，活动，评论】中获取了共计 31 个字段的信息，以半结构化数据的形式存储在 MongoDB 中，每个字段的数据类型和含义如表1所示。

3 数据抽取

原始数据为通过爬虫获取的非结构化数据，需要通过数据抽取将非结构化数据转换为可以直接使用的、剔除无用信息的结构化数据。在数据抽取中对原始数据进行了以下操作：

3.1 数据类型转换

1. 将字符串格式保存的数值类型数据转换为整型或浮点型
2. 将字符串类型保存的日期类型数据转换为时间戳的形式[4]
3. 将布尔类型的数据转换为 0-1 变量

3.2 数据特征抽取

1. 对于文本信息（title, summary text, main text），获取文本长度，并作为新的变量
2. 对于活动信息 (activity)，获取活动数目，并作为新的变量
3. 对于评论信息 (comment_list)，获取评论数目，并作为新的变量
4. 对于图片信息 (img_left_href)，获取图片数目，并作为新的变量
5. 对于标签信息 (tag)，获取标签数目，并作为新的变量

3.3 生成新的特征

1. is_end：通过计算**数据获取时间**与**项目结束时间**之差判断项目是否已结束
2. remain_time：如果项目已结束，则剩余时间为 0，否则剩余时间为**项目结束时间-数据获取时间**
3. conditon：通过 percent 和 is_end 判断项目状态：成功、失败、进行中
4. duration：项目持续时间
5. min_price, max_price, avg_price：**购买选项**中价格的最小值、最大值和均值

[4]Unix 时间戳（Unix timestamp）是一种时间表示方式，定义为从格林威治时间 1970 年 1 月 1 日起至现在的总秒数

表 1: 变量说明

字段	数据类型	含义
id_	int	每个项目的专属 id
collected_money	str	已筹集金额
collected_supporter	str	支持者人数
expiration_date	int	结束时间的时间戳
percent	int	筹集金额与目标金额之比
image_url	str	缩略图地址
title	str	项目标题
url	str	详情页网址
is_new	bool	是否是新项目
is_store_opening	bool	是否可在商店购买
has_target_money	bool	是否有目标金额
has_expiration	bool	是否有截止日期
is_accepting_support	bool	是否接受支持
hide_collected_money	bool	是否隐藏筹集金额
returns	array	–
is_new_store_opening	bool	是否可以在 Makuake STORE 上购买
summary_text	text	摘要部分文字
main_text	text	主要介绍部分文字
target_amount	str	目标金额
thumb_ups	str	facebook 点赞数
activity	Object	所有活动的相关信息
start_at	str	开始日期
end_at	str	截止日期
curr_time	str	当前日期
tag	array	所有标签
category	str	类别
location	str	项目位置
type_	str	项目类型 (all in/all or nothing)
choice_list	array	购买选项详情
img_left_href	array	介绍页面的图片 url
comment_list	array	所有评论

4 数据探索性分析

4.1 数据描述性统计

描述性统计是一种汇总统计，用于定量描述或总结信息集合的特征。通过 df.describe() 生成的描述性统计如表2所示。

表 2: 描述性数据统计

字段	count	mean	std	min	25%	50%	75%	max
id_	29658	15297.69	8723.78	1	7724.25	15256.5	22853.75	30415
collected_money	29658	2401211.73	9344913.74	0	280595	724550	1980617.5	623650600
collected_supporter	29658	217.8	574.41	0	28	79	203	29231
percent	29658	952.37	2600.46	0	108	290	825	104195
title_length	29658	36.98	4.19	5	36	39	40	59
is_new	29658	0	0.07	0	0	0	0	1
is_store_opening	29658	0.13	0.34	0	0	0	0	1
has_target_money	29658	1	0.02	0	1	1	1	1
has_expiration	29658	1	0.01	0	1	1	1	1
is_accepting_support	29658	0.03	0.18	0	0	0	0	1
hide_collected_money	29658	0	0.01	0	0	0	0	1
is_new_store_opening	29658	0.13	0.34	0	0	0	0	1
summary_text_length	29658	89.24	46.94	0	75	105.5	124	432
main_text_length	29658	3556.49	1723.84	4	2370	3267	4408	20649
target_amount	29658	464409.5	1270125.98	0	100000	300000	500000	79600000
thumb_ups	29658	240.92	539.87	0	5	52	235	9427
activity_num	29658	9.59	9.9	0	3	7	13	228
comment_num	29658	27.82	74.34	0	4	11	28	5156
start_at	29658	1.6E+09	6.06E+07	1.38E+09	1.57E+09	1.62E+09	1.64E+09	1.67E+09
end_at	29658	1.6E+09	6.02E+07	1.38E+09	1.58E+09	1.62E+09	1.65E+09	1.68E+09
curr_time	29658	1.67E+09	135158	1.67E+09	1.67E+09	1.67E+09	1.67E+09	1.67E+09
is_end	29658	0.97	0.17	0	1	1	1	1
remain_time	29658	75169.13	503535.09	0	0	0	0	7268658
duration	29658	4.00E+06	2.36E+06	-2.60E+08	2.71E+06	3.74E+06	5.12E+06	2.50E+07
tag_num	29658	8.51	1.21	4	9	9	9	12
choice_num	29658	8.03	4.79	1	5	7	10	106
min_price	29658	9378.47	22317.05	100	2680	4780	9900	1500000
max_price	29658	86633.05	321840.8	500	12411.5	24800	55200	10000000
avg_price	29658	27721.26	61186.59	500	7535.42	13600	26555.56	2694428.57
img_num	29658	29.88	17.13	0	17	27	39	274

从中我们可以看到，collected_money，collected_supporter，percent，target_amount，thumb_ups，comment_num，min_price，max_price，avg_price 的分布都有较为严重的右偏。

几乎所有的商品都不是新商品（is_new），几乎所有商品都有目标金额（target_money）和截止日期（expiration），几乎所有的商品都没有隐藏筹集金额（collected_money）。另外，有 13% 的商品开放商店，有 3% 的商品接受支持。

4.2 数据可视化

核密度图：不同状态项目的变量分布差异

通过核密度图，我们可以更清晰地看到数据的分布。从图6中可以看出，除了描述性统计中提到的几个极度右偏的数据，choice_num，thumb_ups，activity_num 也有较为严重的右偏现象。此外，start_at，end_at，有左偏的现象，表明随着时间发展，网站的项目和活跃用户越来越多。summary_text 出现双峰分布，有大量的数据集中在 0 和 100 附近，tag_num 有大量数据集中在 9 个和 8 个。

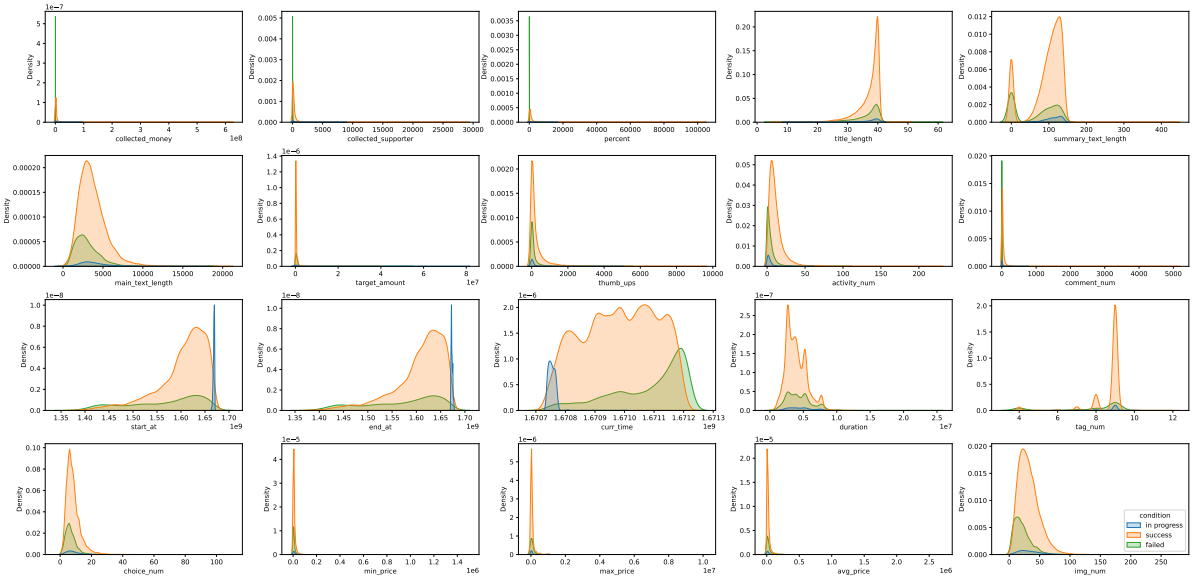


图 6: 不同状态项目的分布差异

箱线图：不同状态项目的变量分布差异

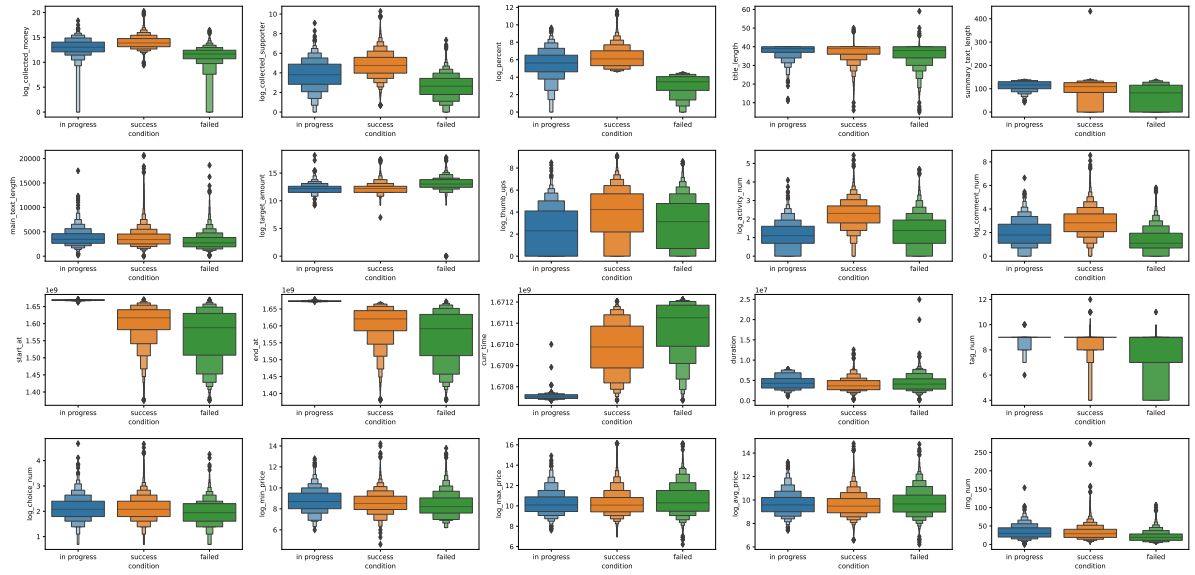


图 7: 不同状态项目的均值及分布差异

随后通过箱线图（如图7所示）来更详细地观察不同状态的项目的均值和分布的差异。在箱线图中，对于前文提到的右偏数据进行了对数处理，以减少数据的偏态。

通过箱线图，可以发现，相较于失败的项目而言，成功的项目在均值和分布意义上都具有更长的标题，更长的摘要，更长的介绍，更多的活动，更多的标签，更高的最低价格，更多的选择以及更多的图片。具有更低的目标金额，更短的持续时间，更低的最低价格以及更低的平均价格。

条形图：类别变量中不同状态项目所占的比例

通过条形图，可以展示成功、失败、进行中的比例与不同类别变量的关系，如图8所示。从中我们可以看到技术类（テクノロジー），化妆品美容类（コスメ・ビューティー），产品类（プロダクト），时尚类（ファッション），食物类（フード）以及餐厅酒吧类（レストラン・バー）有较高的成功率。在爱媛、秋田、富山等地发布的项目有较高的成功率。all in 类型的的项目有较高的成功率。另外，由于只有众筹成功的项目可以在 makuake 中进行贩卖，所以“贩卖中”的成功率为 1。

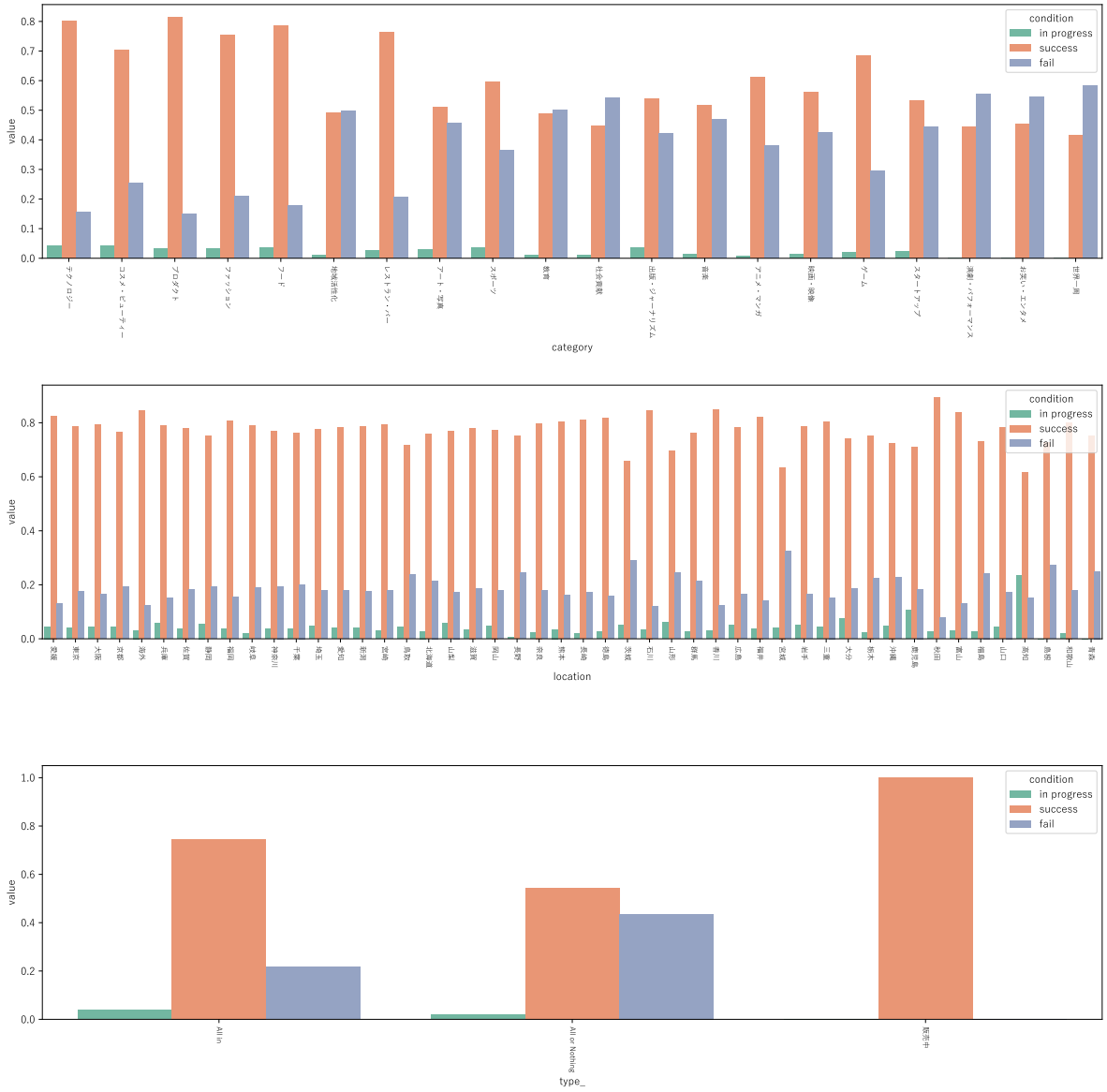


图 8: 类别变量中不同状态项目所占的比例

相关系数矩阵

相关系数矩阵如图9所示。显然，筹集金额与目标金额值比（percent）与筹集金额，支持者人数，点赞数，评论数，标题长度，图片数量，活动数量正相关，与目标金额负相关。

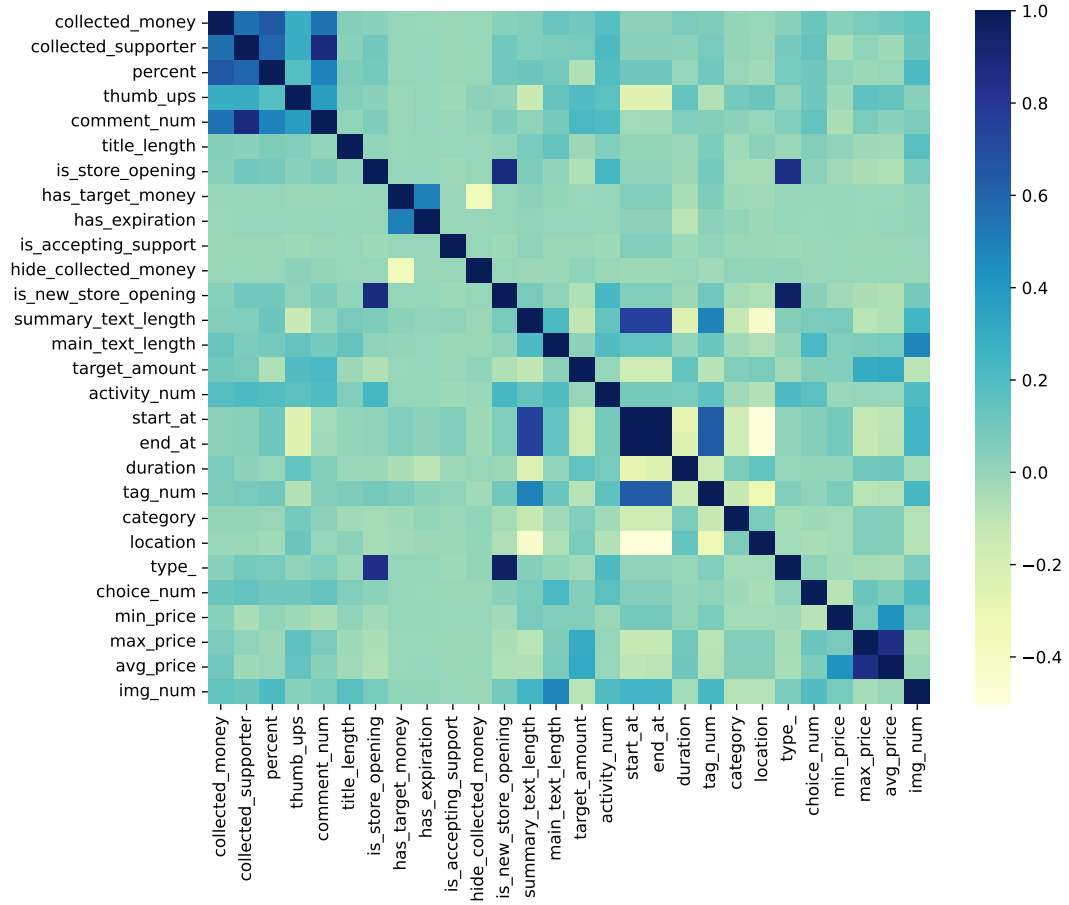


图 9: 相关系数矩阵

5 模型一：利用 OLS 分析众筹项目完成度的影响因素

普通最小二乘法（OLS）是一种根据被解释变量的所有观测值与估计值之差的平方和最小的原则求得参数估计量的方法。最小二满足弗里施 · 沃定理（Frisch Waugh theorem），即解释变量 x_1 的估计系数代表了 x_1 剔除了其它解释变量相关的部分对被解释变量 y 的影响。因此可以单独衡量某个解释变量对于被解释变量的影响。

项目完成度定义为筹款金额与目标金额之比，即：

$$\text{percent} = \frac{\text{collected\_money}}{\text{target\_amount}}$$

5.1 变量选取

由于在已结束项目中，is_end 均为 1，remain_time 均为 0，故剔除这 is_end 和 remain_time 两个变量。由于 duration = end_at - start_at，为线性关系，故剔除 end_at。最后将 tag，category，type_ 三个类别变量转换为哑变量。collected_money，target_amount，activity_num，min_price，max_price，avg_price 6 个变量

进行了对数处理。最终得到 93 个变量。由于研究的是完成度的影响因素，因此并未剔除筹集金额、点赞数、支持者人数等在项目发布时并未得知的变量。

5.2 回归结果

回归结果如表3所示。从中可以看到，筹集金额（collected_money）、支持者人数（collected_suporter）、评

表 3: 回归结果

变量	coef	变量	coef	变量	coef
const	-68.9091***	ファッション	-6.7353**	島根	0.4984
collected_money	1.8981***	フード	-4.5698*	広島	-2.8166**
collected_supporter	0.0283***	プロダクト	-4.6163*	徳島	-2.5706
thumb_ups	0.002***	レストラン　バー	-4.4356	愛媛	-0.4509
comment_num	-0.0299***	世界一周	-3.2019	愛知	-0.3965
title_length	0.0289	出版　ジャーナリズム	-5.0591	新潟	-0.2655
is_store_opening	-1.4346**	地域活性化	-2.6483	東京	0.7819*
has_target_money	85.1172***	教育	-3.4429	栃木	0.55
has_expiration	17.3102	映画　映像	-1.4503	沖縄	-0.3018
is_accepting_support	-1.9068	演劇　パフォーマンス	-1.7738	海外	1.1501**
hide_collected_money	50.2516***	社会貢献	-2.6114	滋賀	1.9262
is_new_store_opening	-4.3593	音楽	-3.0526	熊本	-0.3406
summary_text_length	-0.0035	京都	0.074	石川	-2.863*
main_text_length	3.25E-05	佐賀	-1.2703	神奈川	0.2407
target_amount	-8.0092***	兵庫	-0.3067	福井	-0.9019
activity_num	-0.2013	北海道	-0.8359	福岡	0.9936
start_at	3.02E-09	千葉	-0.1602	福島	3.174
duration	5.62E-07***	和歌山	-3.1825**	秋田	-2.0288
tag_num	-0.7564***	埼玉	1.5824*	群馬	-1.9678
choice_num	0.1152***	大分	-1.103	茨城	-3.2234
min_price	0.417**	大阪	-0.5965	長崎	-2.507
max_price	-4.2687***	奈良	-1.3047	長野	-1.1228
avg_price	8.8882***	宮城	-0.7251	青森	-0.1758
img_num	0.0695***	宮崎	-2.8677	静岡	-0.5138
アニメ　マンガ	-6.245**	富山	-1.9041	香川	-2.3568
アート　写真	-3.2746	山口	-0.8265	高知	-1.1573
ゲーム	-4.7568	山形	-0.4824	鳥取	0.5406
コスメ　ビューティー	-3.9703	山梨	-1.1832	鹿児島	-0.2419
スタートアップ	-3.6069	岐阜	-1.0377	type__All or nothing	2.2701
スポーツ	-3.6863	岡山	-1.1239	type__others	50.2516***
テクノロジー	0.3499	岩手	-0.0118	type__ 販売中	5.8977

论数（comment_num）、是否有目标金额（has_target_money）、是否隐藏已筹集金额（hide_collected_money）、目标金额（target_amount）、持续时间（duration）、标签数量（tag_num）、可选择支持方式的数量（choice_num）、众筹最大金额（max_price）、众筹平均金额（avg_price）、众筹类型 All or Nothing（type__All or Nothing）对项目完成度有着显著影响。

其中，筹集金额、支持人数、点赞数、持续时间、选择数量、平均价格、图片数量、筹集类型中的 All

or Nothing 对完成度有着正向影响。筹集金额每增加 1%，完成度提高 0.018；每增加一个支持人数或一个点赞完成度分别提高 0.028 和 0.002；持续时间每增加一天，完成度上升 0.04；每多一个选择数量，完成度上升 0.11；平均价格每高 1%，完成度上升 0.088；每多一张图片，完成度上升 0.06。

评论数、目标金额、最大众筹金额对完成度有负面影响。值得注意的是每增加一个评论，完成度下降 0.03；目标金额每增加 1%，完成度下降 0.08；最大众筹金额每增加 1%，完成度下降 0.04。

5.3 结论

因此，对于项目发起者来说，想要提高项目成功率，可以选择增加持续时间、给投资者更多的投资选择、增加在 Facebook、Twitter 等公开社交网络上的宣传、增加众筹选项的平均价格、增加介绍图文的图片数量，尝试 All or Nothing 而非 All in 的众筹类型。或者减少目标金额、降低最大众筹金额。

6 模型二：利用机器学习模型预测众筹项目的成功率

除了项目完成度的影响因素，我们更想在项目发布之前或发布之初就能够对该项目是否能够成功进行一个判断。在这一部分，将分别通过逻辑回归和 Lasso 来对众筹项目成功率进行预测。

6.1 变量选择和数据预处理

1. 由于预测是在项目发布之初,因此一些随着项目进行才能得到的变量,如筹集金额（collected_money）、支持者人数（collected_supporter）、完成率（percent）、商店是否打开（is_store_opening）、新商店是否打开（is_new_store_opening）将会被去除。
2. 在已结束项目中完全一致的变量，如是否结束（is_end），剩余时间（remain_time）和状态（condition）也会被去除。
3. 由于 duration = end_at - start_at，为线性关系，故剔除 end_at。
4. 与模型无关的爬取时间（curr_time）也将被剔除。
5. 去除所有文本，即标题（title）、摘要（summary_text）、项目介绍（main_text）。
6. 将所有类别变量转换为哑变量。

最终获得 84 个变量。

6.2 数据预处理

由于目标金额（target_amount）、最小众筹金额（min_price）、最大众筹金额（max_price）和平均众筹金额（avg_price）存在右偏分布的问题，因此我们对这四个变量进行了对数化处理。此外，为了保持量纲的一致性，我们对所有变量进行了 Z-score 标准化处理。

6.3 逻辑回归

逻辑回归是一种分类方法。将线性回归在 $\mathbb{R}$ 上的结果通过 sigmoid 函数映射到 $[-1, 1]$ 上，以表示事件发生的概率。其模型为：

$$h(x) = \sigma(\mathbf{w}^\top \mathbf{x})$$

6.3.1 回归结果

逻辑回归的回归结果如图10所示。其中每个方格对应一个权重的系数,红色表示系数为正,蓝色表示系数为负。为了能够显示更多的颜色,超出 1 的权重以 1 来表示。可以看到,是否有目标金额（has_target_money）、

摘要长度（summary_text_length）、目标金额（target_amount）、活动数量（activity_num）、开始时间（start_at）、持续时间（duration）、标签数量（tag_num）、选择数量（choice_num）、最大众筹价格（max_price）、图片数量（img_num）、是否为食品类（category_ フード）、是否为产品类（category_ プロダクト）、是否为餐厅酒吧（category_ レストラン　バー）对于预测有着比较强的影响。

图 10: 逻辑回归权重示意图

6.3.2 模型评估

逻辑回归的训练集精度为 0.85297，测试集精度为 0.85254。测试集分类报告如表4所示。相比于随机猜（约 78.5%）提高了约 6.7 个百分比。

表 4: 逻辑回归测试集分类报告

	precision	recall	f1-score	support
0.0	0.74	0.47	0.58	1216
1.0	0.87	0.95	0.91	4528
accuracy			0.85	5744
macro avg	0.80	0.71	0.74	5744
weighted avg	0.84	0.85	0.84	5744

6.4 Lasso

Lasso 是一种回归方法，在线性回归的基础上加入 L1 正则项以进行变量筛选。相比于线性回归，Lasso 不容易过拟合，在测试集上有着较好的表现。其优化函数如下：

$$h(x) = \frac{1}{2}||A\mathbf{x} - b||^2 + \lambda||\mathbf{x}||_1$$

在众筹问题中，可以先对于完成度进行回归拟合，再根据拟合结果是否大于 100% 来判断该项目是否成功。相比于逻辑回归而言，Lasso 虽然不能表明有多大概率成功，但是可以表明可以达到怎样的一个完成度。是以极高的完成率成功，还是以较低的完成率成功。

6.4.1 回归结果

对于不同的 λ 取值，Lasso 筛选变量的效果效果和精度也有所不同。图11展示了随 λ 变化，解的变化情况。可以看到，随着惩罚项系数的增加，有越来越多的变量系数趋近并等于 0。

图12展示了随 λ 变化，训练集精度的变化情况。可以看到，随着惩罚项系数的增加，训练集精度呈现先上升后下降的趋势。

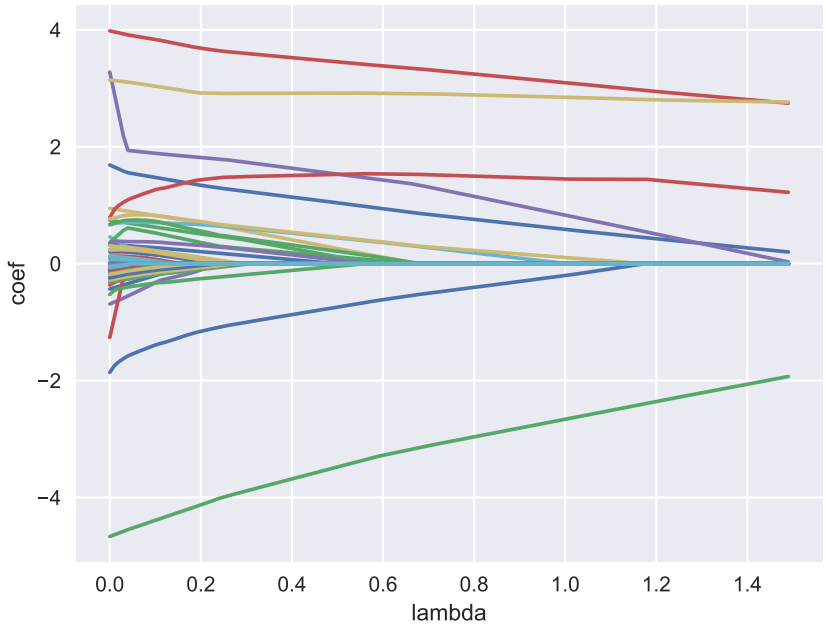


图 11: Lasso 解的路径

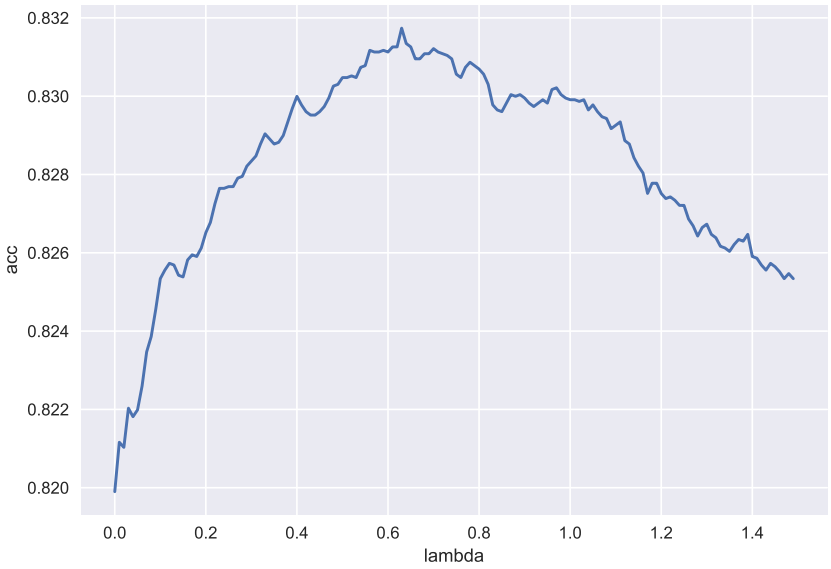


图 12: Lasso 在训练集上的精度随 λ 的变化

以精度最高的 0.6 作为惩罚项系数，进行回归。其回归结果如图13所示。其中每个方格对应一个权重的系数，红色表示系数为正，蓝色表示系数为负。为了能够显示更多的颜色，超出 1 的权重以 1 来表示。可以看到，除了目标金额（target_amount）、活动数量（activity_num）、持续时间（duration）选择数量（choice_num）、最低众筹价格（min_price）、平均众筹价格（avg_price）、图片数量（img_num）、是否为技术类（category_ テクノロジー）、是否为时尚类（category_ ファッション）、是否为产品类（category_ プロダクト）、是否为东京（location_ 東京）和海外（location_ 海外）以外的变量的系数均被压缩为了 0。

图 13: Lasso 权重示意图

6.4.2 模型评价

Lasso 回归的训练集精度为 0.83113，测试集精度为 0.83269。测试集分类报告如表5所示。相比于随机猜（约 78.5%）提高了约 4.7 个百分点。

表 5: Lasso 测试集分类报告

	precision	recall	f1-score	support
0.0	0.68	0.40	0.50	1230
1.0	0.85	0.95	0.90	4514
accuracy			0.83	5744
macro avg	0.77	0.67	0.70	5744
weighted avg	0.82	0.83	0.82	5744

7 模型三：利用深度学习模型预测众筹项目的成功率

在模型二中，主要利用了数字特征对众筹项目的成功率进行预测。但是在实际生活中，标题可以吸引人们的兴趣，图片和文字则是人们用来判断是否要支持某一项目的重要依据。因此标题和介绍这类文本特征显然也是影响成功率的一个主要因素。在这一节中，我们希望能够通过引入文本特征来加强成功率预测的准确率。

7.1 文本数据预处理

7.1.1 分词

文本分词，也称为 Tokenization，是将文本分解为较小的单元（称为 Token）的过程，这些单元可以是单词、短语、符号或其他元素。在自然语言处理中，文本分词非常重要，因为它允许文本以数值格式

表示，进而作为机器学习算法的输入。分词也有助于减小词汇表的大小，从而使模型更有效地处理数据。此外，分词还可以从文本中识别和提取有意义的信息，例如命名实体、词性等等。

在本次实验中，我们采用了 Transformer[5] 中的 tokenizer 来对文本进行分词。tokenizer 的优点在于其使用了 Subword tokenization 的分词方式，这种分词方式能够有效地降低字典大小，减小模型参数。以“Let's do tokenization!”为例，以空格、标点、字符和子词（subword）四种方式分词的结果如图14所示。

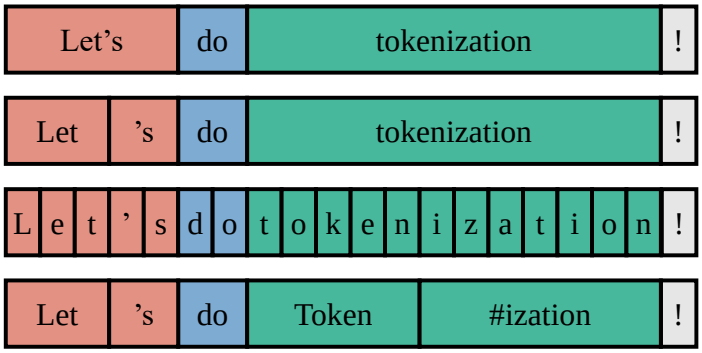


图 14: 示例：四种分词方式

tokenization 与 customization、standardization、modernization、organization 等词共用了 #ization 的词根，与 tokenize、tokenizer、tokenizing 等词共用了 token 的前缀。从而大大减少了字典大小。并且在遇到语料库中不存在的词语时，Subword tokenization 的分词方式还可以通过将陌生词切分成已知的子词从而使得模型获得处理它之前没见过的词的能力。

7.1.2 文本数据清洗

文本数据清洗是为进一步分析或建模准备原始文本数据的过程。文本数据清洗的目的是删除或纠正文本中任何不相关、不正确或不一致的信息，例如拼写错误、标点错误和 HTML 标签，使数据格式适合进一步的分析或建模。

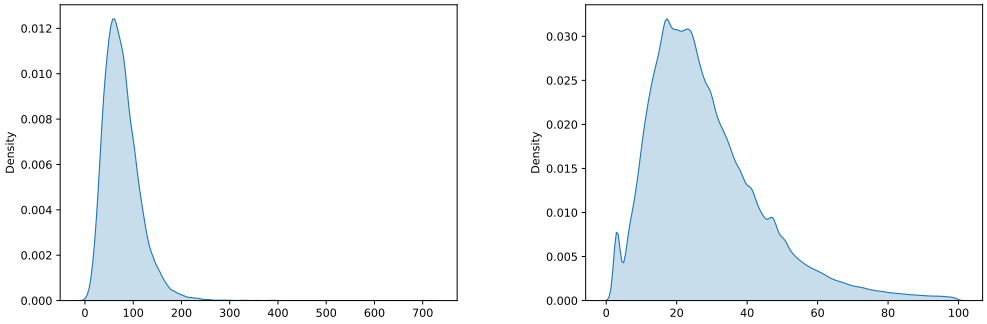
在本次实验中，我们首先去除了\xa0、\u3000 和空格这三种常见的空白间隔符，其次去除了包括引号、书名号、尖括号、方括号在内的各种中英文标点符号和特殊字符。最后，以句号、叹号、问号、换行符为分隔符，对文本进行分句。值得注意的是，仅按句子层面的模型需要进行分句。

7.1.3 Padding and Truncate

Padding 和 Truncate 是自然语言处理中的两种常见处理方式。用于解决不同文本长度导致的输入与模型维度不一致的问题。Padding 是在短文本的末尾补充特殊字符使得所有文本长度相同，而 Truncate 是将长文本截取一部分内容使得所有文本长度相同。

绘制句子数量分布图和词语数量分布图，如图15所示。可以发现，一篇文章的句子数量大多在 120-150 以内，一个句子中的词语数量大多在 50 以内。一篇文章的词语大多在 2500-3000 以内。因此，对于以句子为单位的模型，需要将一篇文章分别在词语层级和句子层级进行 padding 和 truncate，将文章转换为（sentence_per_document，word_per_sentence）的矩阵。而对于以文章问单位的模型，只需将文本转换为长度为 word_per_document 的向量即可。

[5]Transformer: 由 Hugging Face 团队维护，是一个开源的自然语言处理模型库，提供了大量基于 Transformer 模型的预训练模型。



(a) 文本中句子数量　　(b) 句子中词语数量

图 15: 寻找合适的文本长度和句子长度.

7.2 数值数据预处理

数值预处理部分和模型二中一致，首先删除了与预测成功率无关的变量，如已筹集金额、爬取时间等等，其次将类别变量转化为哑变量。最后对存在右偏分布的变量进行对数化处理，再对所变量进行标准化处理，具体操作如如6.1节6.2节所示。

7.3 构建 DataLoader

Dataset 和 DataLoader 是 PyTorch 中构建训练数据的重要模块。Dataset 是一个抽象数据集接口，可以通过索引返回单条数据。DataLoader 是一个用于加载数据的类，可以从 Dataset 中读取数据，并按 batch_size 加载数据，以解决内存不足问题；对数据进行打乱，从而避免过拟合问题。

在常见的图片分类等任务中，往往都是在 Dataset 中存入图片的地址，等调用的时候再读取图片并进行预处理。在本次实验中，由于文本数据较小，但是分词时间较长，所以在初始化 Dataset 时便将文本转换为张量，以避免训练过程中的分词，从而提升 GPU 的利用率和训练速度。

另外，由于样本存在数据偏态问题，本次实验中采用了 Sampler 来对数据进行抽样，用正类和负类出现频率的倒数作为采样权重进行抽样，以保证每个 epoch 学到的数据都是近似平衡的数据。其具体过程如图16所示。

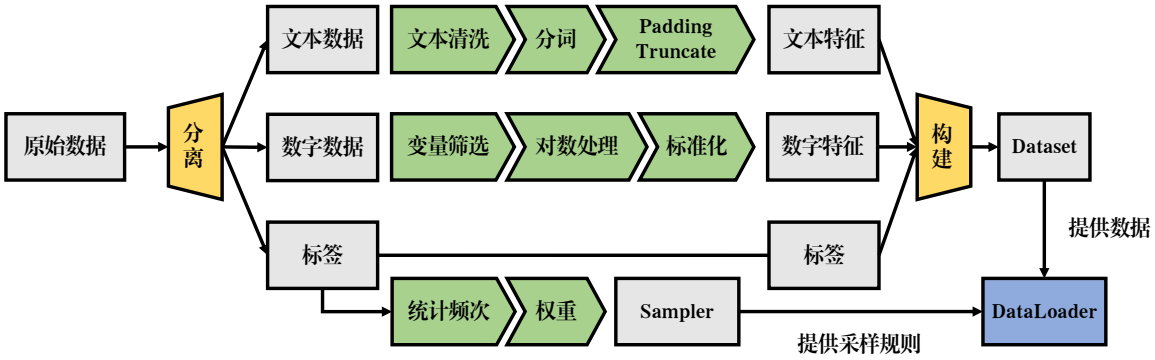


图 16: 构建 DataLoader 框架示意图

7.4 模型框架

模型框架如图17所示。该模型有两个输入，分别为文章和数字特征。文章通过某种模型转换为一个表示文本特征的向量。再通过全连接层与数字特征进行特征融合，得到一个新的特征向量。由于是二分类问题，最后再通过全连接层将特征映射到一个数，通过 sigmoid 函数转换为概率并计算 loss，反向传播并更新权重。

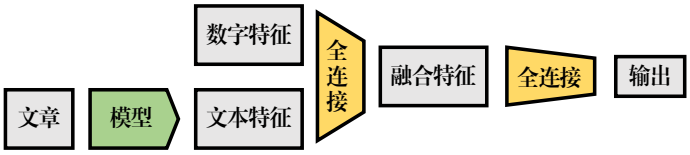


图 17: 深度学习模型框架

7.5 通过 gru 提取文本特征

7.5.1 模型介绍

GRU (Gated Recurrent Unit) 是一种循环神经网络（RNN）的变体。GRU 通过更新门和重置门来更改隐藏状态，从而改进了传统 RNN 的长期记忆问题。

一个双层 GRU 网络如图18所示。每个 GRU Cell 接受两个输入：当前时刻的序列的输入 x_t 和上一时刻的隐藏层 h_{t-1}。经过门运算后，更改隐藏层，并输出 y_t。第二层的 GRU Cell 以第一层的输出作为当前时刻的输入。我们以最后一层的最后时刻的输出做为文本特征。并将该模型添加到7.4节中“模型”对应的位置。

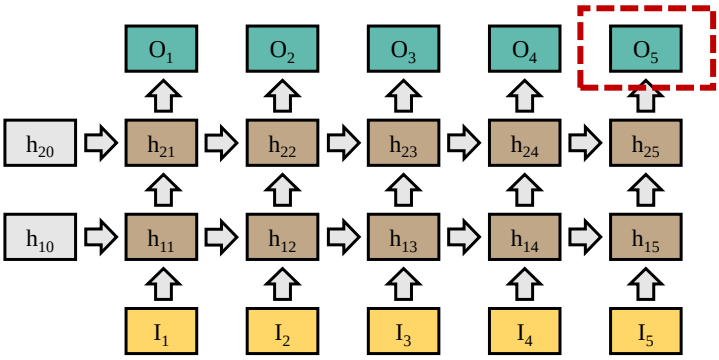


图 18: GRU 网络结构示意图

7.5.2 运行结果

图19展示了词向量长度为 64、GRU 隐藏层大小为 256、全连接层大小为 64 的情况下损失与精度随 epoch 的变化。适当减小模型参数，词向量长度为 16、GRU 隐藏层大小为 32、全连接层大小为 16 的情况下损失与精度随 epoch 的变化如图20所示。

7.5.3 模型评价

基于 GRU 提取文本特征的模型在训练集上取得了很好的效果，在训练集上的准确率一度达到了 0.9936。但是不论如何减小模型参数，该模型在测试集上的表现都不是很好，最高准确率为 0.8466，小于逻辑回

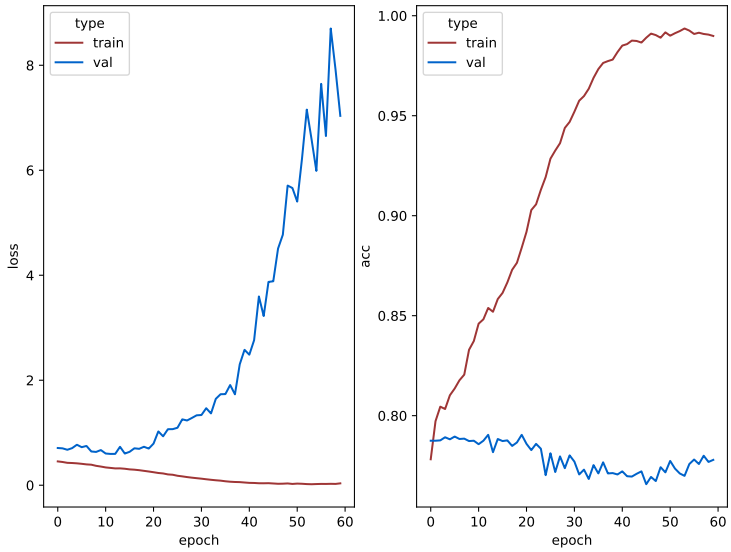


图 19: 基于 GRU 提取文本特征的模型的损失与精度图

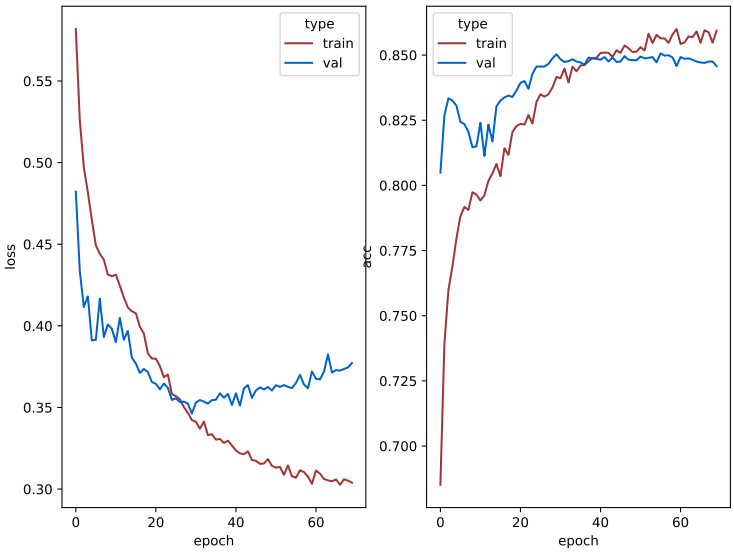


图 20: 基于 GRU 提取文本特征的模型的损失与精度图

归的 0.8525。引入文本特征使得神经网络规模增大，学习能力更强，但是过长的文本导致的梯度消失问题使得模型更倾向于记住训练数据，而难以学到有用的信息。

7.6 通过 HAN 提取文本特征

7.6.1 模型介绍

Hierarchical Attention Network (HAN) 是一种用于文本分类的深度学习模型 (Yang et al., 2016)。整个模型分为五层，具体结构如图21所示。从下往上依次是：词级别的双向 GRU 层、词级别的注意力机制层、句子级别的双向 GRU 层，句子级别的注意力机制层和输出层。注意力机制可以表示为 $value = f(key, Query)$。注意力机制根据模型的输出和一组 key 的运算，得到注意力分数，注意力分数的大小在一定程度上代表该时刻输入的词对于模型的重要程度。

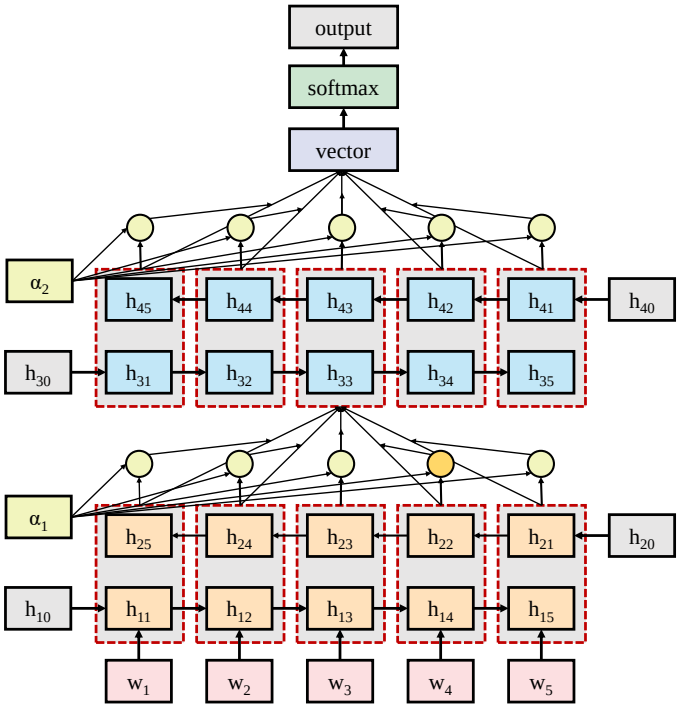


图 21: HAN 网络结构示意图

整个模型的流程如下：首先将每个句子中的词作为输入，输入词级别的双向 GRU。双向 GRU 的输出经过注意力机制得到注意力分数并加权求和后，得到表示句子信息的向量。再将一篇文章的所有表示句子作为输入，输入句子级别的双向 GRU。双向 GRU 的输出经过注意力机制得到注意力分数并加权求和后，得到表示文章信息的向量。最后再经过 softmax 函数得到各个类别的概率。

可以看到，在处理长文本的时候，相比于 GRU，HAN 只需要句子个数 + 句子长度级别的 GRU Cell，大大减少了模型参数，降低了过拟合的风险。

在本次实验中，我们将 HAN 得到的表示文章信息的向量作为文本特征。并将该模型加入到7.4中“模型”的位置。

7.6.2 运行结果

基于 HAN 提取文本特征的模型的损失与精度如图22、图23所示。其中图22在特征融合时采用了合并的（concat）的方法，而图23则使用了加和的方法。加和的方法在精度上更具优势，过拟合也更慢。具体参数如表6所示。在 HAN 模型的代码方面，我参考了sgrvinod的a-PyTorch-Tutorial-to-Text-Classification库。

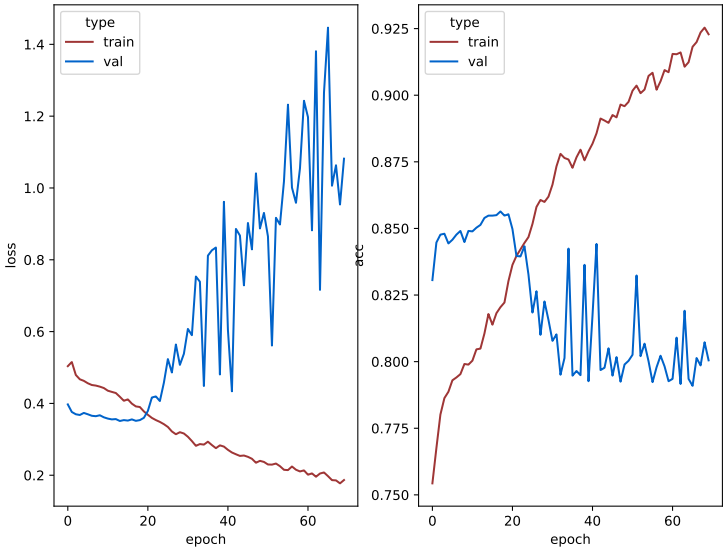


图 22: 基于 HAN 提取文本特征的模型的损失与精度图

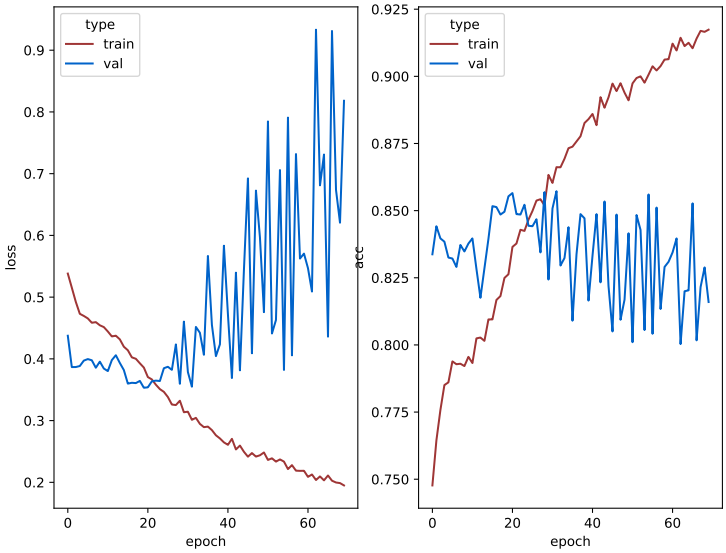


图 23: 基于 HAN 提取文本特征的模型的损失与精度图

7.6.3 模型评价

基于 HAN 提取文本特征的模型在测试集上的精度为 0.8572，比逻辑回归的 0.8525 高出 0.47 个百分点，提升较为有限。

表 6: 基于 HAN 提取文本特征的模型的参数

参数名称	参数大小
epochs	70
batch_size	256
lr	2e-3
embed_size	16
dense_hidden_size	32
word_gru_size	16
word_gru_layers	2
word_att_size	16
sentence_gru_size	16
sentence_gru_layers	2
sentence_att_size	16
word_per_sentence	40
sentence_per_document	120
dropout	0.5

7.7 通过 BERT 提取文本特征

7.7.1 模型介绍

BERT (Bidirectional Encoder Representations from Transformers) 是一种用于自然语言处理的预训练模型 (Devlin et al., 2018)。其具体结构如图24所示。BERT base 由 Encoder 和 12 个 Transformer Encoder 块构成。其中，Encoder 层包含三个维度为 768 的 Embedding 层，分别对应词、位置和句子类型。最后的词向量由三个 Embedding 层的输出相加得到。Transformer Encoder 块由一个包含 12 个自注意力机制的多头注意力机制、一个前馈神经网络和两个标准化层组成。能够处理双向上下文关系，捕捉词语的语境信息。BERT 通过预测掩码位置的词语（Masked LM）和预测下一个句子（Next Sentence Prediction）两个任务来的到预训练权重。在下游任务中，采用预训练权重作为初始权重。BERT 的预训练技术也使得在没有大量标注数据的情况下，仍然可以获得较高的模型性能。

在本次实验中，我们将 BERT 中 <CLS> 对应的输出作为表示句子信息向量，对所有句子级别的向量加和得到表示文章信息的向量。以该向量作为文本特征，并将修改过的 BERT 模型加入到7.4中“模型”的位置。

7.7.2 运行结果

基于 HAN 提取文本特征的模型的损失与精度如图25所示。在模型的代码方面，采用了 transformers 库的 BERTModel 类，在模型权重方面，采用了来自东京大学的bert-small-japanese，相比于 BERT-base，BERT-small 每个多头感知机只有 4 个自注意力机制，相应的，隐藏层也由 768 维下降至 256 维。同时支持的最大句子长度由 512 下降至 128。相对于 BERT-base，BERT-small 具有更小的模型，更快的训练速度，且理论上更不容易过拟合。

7.7.3 模型评价

基于 BERT 提取文本特征的模型在测试集上达到了 0.8593 的精度，比 HAN 高 0.021 个百分点，比逻辑回归高 0.68 个百分点。模型过拟合的速度也相对较慢，后续测试集的损失和精度的波动也在一个较小的范围内。但是无法继续下降。

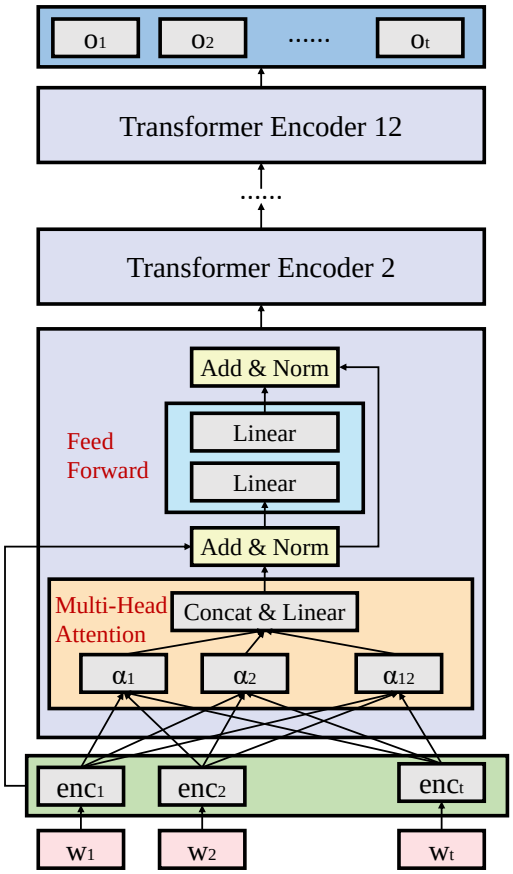


图 24: BERT 网络结构示意图

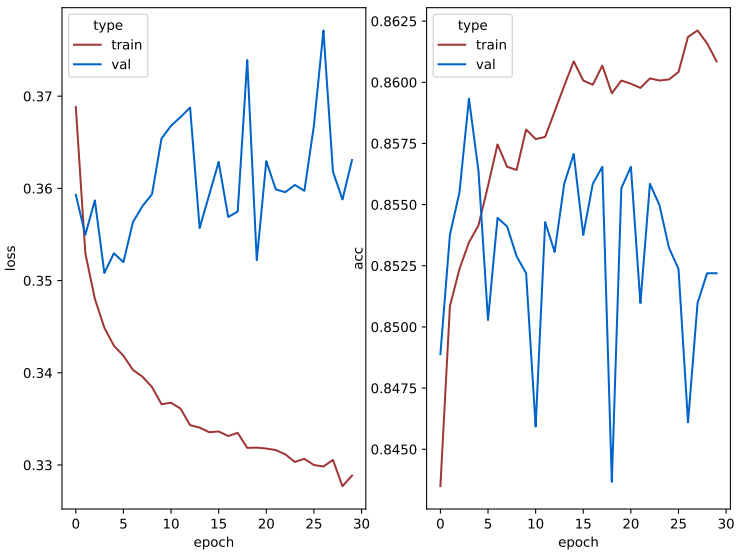


图 25: 基于 BERT 提取文本特征的模型的损失与精度图

7.8 总结

虽然试图通过引入文本特征来增加模型预测的准确率，但是从实际结果而言，基于 GRU 提取文本特征的模型并不能提升模型预测的准确率。而基于 HAN 和 BERT 的模型提升的准确率也有限。根据分析，有可能是基于以下原因：

1. 数据量不足。虽然经过清洗后的数据有将近 2.9 万条，但是这个数据量对于深度学习模型而言还是太少了。即使是按照 Subword tokenization 的分词方式，字典的大小也超过了 3 万。因此在 Embedding 层大约有 52 万个参数，而 BERT 模型由于词向量大小为 256，在 Embedding 层有超过 800 万个参数。而 GRU 部分，HAN 模型有将近 8.2 万的参数，GRU 模型有 50 万以上的参数。BERT 的每个 Transformer Encoder 中，也有 78 万个参数。因此 3 万的数据量对于深度学习的模型是微不足道的，并且 GRU 和 HAN 并没有预训练权重，因此十分容易过拟合。
2. 众筹是否成功的随机性较强。有可能两个十分相似的项目，但是有截然不同的结果。除了从网站上获取到的各种因素外，众筹项目能否成功或许还和在社交网络上是否有过宣传，是否得到了众筹平台的推荐、是否得到了其它平台的推荐、是否赶上了潮流等等。这些数据往往难以获得，或者没有显式的数据可以表达。因此导致预测成功率的精度的上限较低。
3. 训练集与测试集的分布不一致。不同的训练集和测试集的划分也会影响到最终的结果。有可能在文本数据方面，每篇文章之间的差异较大，导致训练集和测试集之间也会有较大的差异。
4. 模型设计问题。由于将文本特征和数字特征部分融合并预测的部分的模型是自己设计的，在设计方面可能存在一些缺陷，使得模型的表现无法继续提升。

8 模型四：基于深度学习模型预测众筹项目类别

最后，利用 HAN 模型对项目类别进行预测，根据预测结果，平台可以在用户项目提交时自动推荐适合类别，从而提高服务质量和用户体验。

由于类别存在着严重的数据不平衡（如图26所示），产品类有 15064，占到了整个数据集的一半以上，其余大多数类别从几百到几千不等，最少的“世界一周”仅有十二条样本。因此依据出现频率的倒数作为权重进行采样，以保证每个 epoch 模型学习到的样本都近似平衡样本。

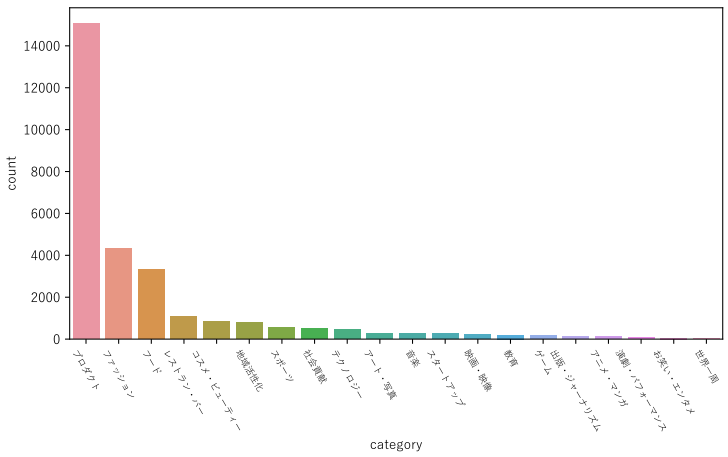


图 26: 项目类别数量统计

8.1 运行结果与模型评价

模型损失与精度随 epoch 的变化如图27所示。可以看到，训练集损失逐步下降、精度逐步上升。100 个 epoch 后，训练集精度达到了 0.92。而测试集方面，在 40 个 epoch 之后开始逐渐出现过拟合现象，loss 逐渐上升。测试集准确率最高达到了 76.74%。介于文本分类准确率普遍不是很高，虽然说有严重的数据不平衡，随机猜的准确率为 56%，但这仍然算是一个不错的准确率。

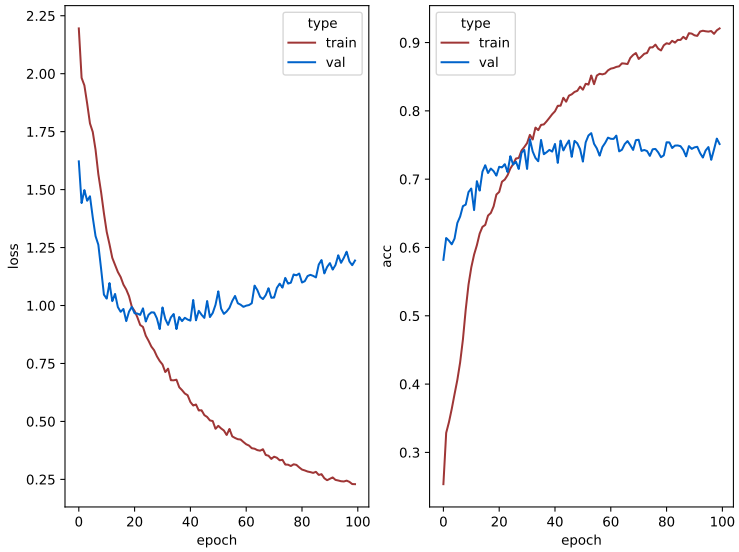


图 27: 基于 HAN 模型预测项目类别的损失和准确率随 loss 的变化

8.2 结果可视化

另外，基于注意力权重的特性，我们可以将其可视化，以清晰地看到哪些文本在模型的预测中显得比较重要。我们将注意力分数大于 0.6 的词语挑出来，依据注意力分数的大小设定不同的字体大小和颜色。其结果如图28-30所示。图片的左上角表明了预测的类别和准确率，句子的左侧的红条表明句子的重要程度，越大的词语表明其注意力分数越高。可以看到，在游戏类别中，模型更注重“革新”、“中级”、“上级”、“世界”、“制作”等词。而在社会贡献中，更关注“地震”、“援助金”、“支援”、“灾害”等词。在食物中，更关注“老铺”、“味噌屋”、“BBQ”等词。

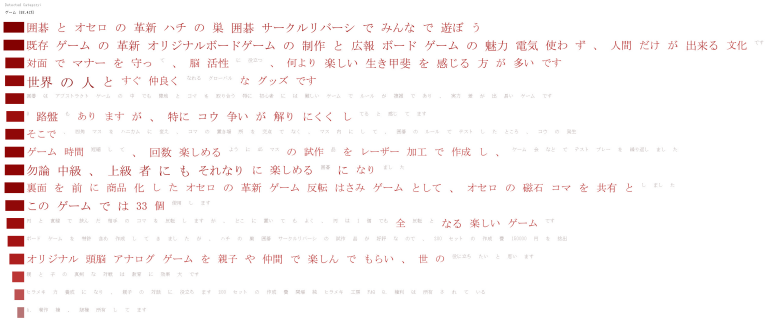


图 28: 游戏类别中权重较大的词语的可视化

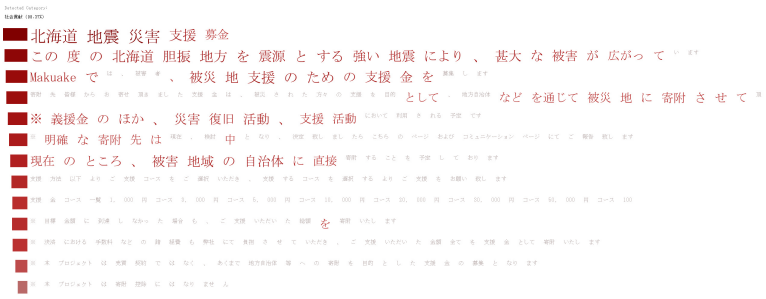


图 29: 社会贡献类别中权重较大的词语的可视化

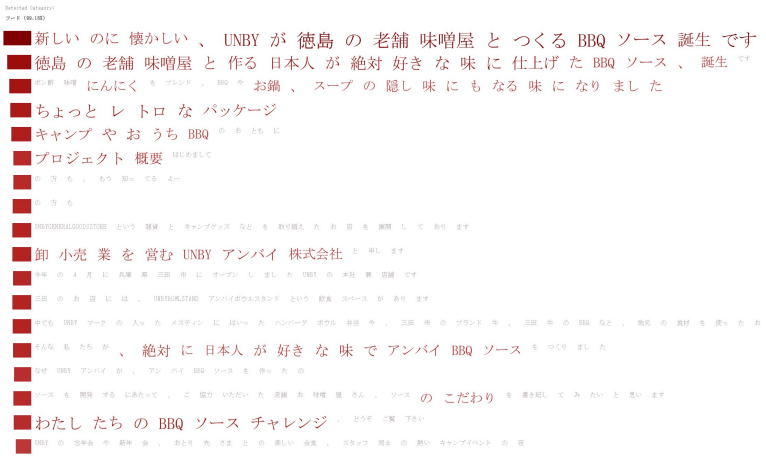


图 30: 事物类别中权重较大的词语的可视化

9 结论与展望

9.1 结论

众筹项目完成度方面，基于 OLS 可以得出：筹集金额、支持者人数、评论数、是否有目标金额、是否隐藏已筹集金额、目标金额、持续时间、标签数量、可选择支持方式的数量、众筹最大金额、众筹平均金额、众筹类型 All or Nothing 对项目完成度有着显著影响。其中，筹集金额、支持人数、点赞数、持续时间、选择数量、平均价格、图片数量、筹集类型中的 All or Nothing 对完成度有着正向影响。评论数、目标金额、最大众筹金额对完成度有负面影响。目标金额对于完成度的影响最大。

在成功率预测方面，效果的优劣依次为：BERT>HAN>Logistic Regression>GRU>Lasso，所有的深度学习模型都出现了过拟合的现象，文本特征对于成功率的提升有限。

在文本分类方面，模型准确率最高达到了 76.74%，效果不错。注意力机制可视化也可以帮助我们加深对模型运作模式的理解。

9.2 展望

Makuake 中的全部项目不到 3 万条，数据较少，但是如Kickstarter、Indiegogo等众筹网站，虽然数据量较大，但是只能获取到按照平台推荐排序的项目，数据存在较为严重的不平衡。且目前已有的 Kickstarter 或 Indiegogo 的数据集并未包含文本特征。希望以后能够拓展数据集的大小和内容，从而构建出更加完善的模型。

本文只是利用三个全连接层对数据特征和文本特征进行了融合和预测。未来可以探索更多更有效的融合方式。

视频和图片信息对于众筹成功率也有着显著的影响。未来可以引入图像特征，和数字特征与文本特征相融合，从而可以更好地模拟人在做决策时所接受到的信息。

参考文献

[1] Ajay Agrawal, Christian Catalini, and Avi Goldfarb. “Some simple economics of crowdfunding”. In: *Innovation policy and the economy* 14.1 (2014), pp. 63–97.

[2] Paul Belleflamme, Thomas Lambert, and Armin Schwienbacher. “Individual crowdfunding practices”. In: *Venture Capital* 15.4 (2013), pp. 313–333.

[3] Valerie Busse and Michal Gregus. “Crowdfunding – An Innovative Corporate Finance Method and Its Decision-Making Steps”. In: *Advances in Intelligent Networking and Collaborative Systems*. Ed. by Leonard Barolli, Hiroaki Nishino, and Hiroyoshi Miwa. Cham: Springer International Publishing, 2020, pp. 544–555. ISBN: 978-3-030-29035-1.

[4] Douglas J. Cumming, Gaël Leboeuf, and Armin Schwienbacher. “Crowdfunding models: Keep-It-All vs. All-Or-Nothing”. In: *Financial Management* 49.2 (2020), pp. 331–360. DOI: https://doi.org/10.1111/fima.12262. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/fima.12262. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/fima.12262.

[5] Jacob Devlin et al. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805* (2018).

[6] Vincent Etter, Matthias Grossglauser, and Patrick Thiran. “Launch hard or go home! Predicting the success of Kickstarter campaigns”. In: *Proceedings of the first ACM conference on Online social networks*. 2013, pp. 177–182.

[7] Jirka Härkönen. “Crowdfunding and its utilization for startup finance in Finland ⁻factors of a successful campaign”. In: *Lutpub* (2014).

[8] Anna Lukkarinen et al. “Success drivers of online equity crowdfunding campaigns”. In: *Decision Support Systems* 87 (2016), pp. 26–38. ISSN: 0167-9236. DOI: https://doi.org/10.1016/j.dss.2016.04.006. URL: https://www.sciencedirect.com/science/article/pii/S0167923616300598.

[9] Ethan Mollick. “The dynamics of crowdfunding: An exploratory study”. In: *Journal of Business Venturing* 29.1 (2014), pp. 1–16. ISSN: 0883-9026. DOI: https://doi.org/10.1016/j.jbusvent.2013.06.005. URL: https://www.sciencedirect.com/science/article/pii/S088390261300058X.

[10] Zichao Yang et al. “Hierarchical attention networks for document classification”. In: *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*. 2016, pp. 1480–1489.

[11] Haichao Zheng et al. “The role of multidimensional social capital in crowdfunding: A comparative study in China and US”. In: *Information & Management* 51.4 (2014), pp. 488–496. ISSN: 0378-7206. DOI: https://doi.org/10.1016/j.im.2014.03.003. URL: https://www.sciencedirect.com/science/article/pii/S0378720614000305.