



北京大学

Python Data Analysis Report

Analysis and Prediction of the Success Rate of Startups

Xingjian Liu, Wonduk Seo, Shutian Liang

二〇二二年六月

目录

| | |
|--------------------------|----|
| 一、选题背景 | 1 |
| 1. 1. 研究背景 | 1 |
| 1. 2. 研究内容 | 2 |
| 1. 3. 研究意义 | 2 |
| 二、成功企业画像 | 3 |
| 2. 1. 数据介绍 | 3 |
| 2. 1. 1. 数据来源 | 3 |
| 2. 1. 2. 数据检视 | 3 |
| 2. 2. 企业基础信息画像 | 5 |
| 2. 2. 1. 公司运营情况 | 5 |
| 2. 2. 2. 公司的时空分布 | 6 |
| 2. 2. 3. 公司的产业类型分布 | 7 |
| 2. 3. 创业者画像 | 9 |
| 2. 3. 1 创业者的工作经历 | 9 |
| 2. 3. 2 创业者的受教育程度 | 10 |
| 2. 3. 3 创业者的技能得分情况 | 12 |
| 2. 4. 创业公司画像 | 13 |
| 2. 4. 1. 创业公司的规模 | 13 |
| 2. 4. 2. 创业公司的团队组成 | 14 |
| 2. 4. 3 创业公司的最高学历 | 15 |
| 2. 4. 4. 创业公司的福利机制 | 17 |
| 2. 5. 小结 | 18 |
| 2. 5. 1. 结果总结 | 18 |
| 2. 5. 2. 相关建议 | 18 |
| 三、企业成功率预测模型 | 19 |
| 3. 1. 数据介绍 | 19 |

| | |
|----------------------------------|----|
| 3.1.1. 数据来源 | 19 |
| 3.1.2. 数据检视 | 19 |
| 3.1.2.1 特征字段详情 | 19 |
| 3.1.2.2 缺失值情况 | 21 |
| 3.2. 探索式数据分析 | 22 |
| 3.2.1. 企业的空间分布 | 22 |
| 3.2.2. 企业的时间分布 | 23 |
| 3.2.3. 企业的产业类别分布 | 23 |
| 3.2.4. 企业的投资人情况 | 24 |
| 3.2.5. 企业的集资情况 | 26 |
| 3.2.6. 企业与外界的关系网 | 28 |
| 3.3. 数据预处理 | 29 |
| 3.3.1. 删除无意义字段 | 29 |
| 3.3.2. 删除质量较差的字段 | 29 |
| 3.3.3. 处理 category_code 字段 | 30 |
| 3.4. 机器学习模型构建 | 31 |
| 3.4.1. 选取模型 | 31 |
| 3.4.2. 调节超参数 | 33 |
| 3.4.3. 建立集成模型 | 36 |
| 四、总结及反思 | 40 |

图目录

| | |
|-----------------------------------|----|
| 图 1 研究方向思维导图..... | 2 |
| 图 2 全部特征字段分类 | 3 |
| 图 3 有效字段筛选..... | 4 |
| 图 4 该数据集中成功与失败企业占比情况..... | 5 |
| 图 5 数据集中不同公司成立年份分布..... | 6 |
| 图 6 数据集中不同公司的地区分布..... | 6 |
| 图 7 成功企业产业类别词云图..... | 8 |
| 图 8 失败企业产业类别词云图..... | 8 |
| 图 9 创业者工作经历堆积图..... | 9 |
| 图 10 创业者毕业学校堆积图..... | 10 |
| 图 11 两类公司中创业者毕业学校堆积图..... | 11 |
| 图 12 成功企业与失败企业创业者技能得分的密度分布图..... | 12 |
| 图 13 成功企业与失败企业创业者技能得分的箱线图..... | 12 |
| 图 14 成功企业与失败企业的公司规模..... | 13 |
| 图 15 成功公司与失败公司结构评分情况..... | 14 |
| 图 16 成功公司与失败公司创业团队中最高学历分布饼图..... | 15 |
| 图 17 成功公司与失败公司创业团队中最高学历分布雷达图..... | 15 |
| 图 18 成功公司与失败公司创业团队中人才类型分布饼图..... | 16 |
| 图 19 成功公司与失败公司创业团队中薪资结构评价..... | 17 |
| 图 20 成功公司与失败公司创业团队中的奖励机制..... | 17 |
| 图 21 机器学习数据集缺失值情况分布图..... | 21 |
| 图 22 机器学习数据集企业地理位置分布等值域图..... | 22 |
| 图 23 机器学习数据集企业时间分布折线图..... | 23 |
| 图 24 机器学习数据集企业产业类型分布..... | 23 |
| 图 25 企业是否有风险投资人柱状图..... | 24 |
| 图 26 企业是否有天使投资人柱状图..... | 25 |
| 图 27 平均每轮参与集资的投资者数量密度分布图..... | 25 |
| 图 28 平均每轮参与集资的投资者数量密度分布图..... | 26 |

| | |
|---------------------------------|----|
| 图 29 企业集资轮数分布图..... | 27 |
| 图 30 企业每轮集资是否进行分布图..... | 27 |
| 图 31 企业与外界关系网数量分布图..... | 28 |
| 图 32 预处理后数据集缺失值分布..... | 29 |
| 图 33 机器学习各模型准确率比较..... | 31 |
| 图 34 机器学习各模型交叉验证平均准确率比较..... | 32 |
| 图 35 贝叶斯调参过程..... | 33 |
| 图 36 Optuna 调参过程..... | 34 |
| 图 37 调参后模型与原模型做交叉验证平均准确率比较..... | 35 |
| 图 38 集成模型的混淆矩阵图..... | 36 |
| 图 39 所有模型 ROC 曲线比较..... | 38 |

表目录

| | |
|--------------------------|----|
| 表 1 不同类别学校毕业的时候比较列表..... | 10 |
| 表 2 机器学习数据集特征字段详情列表..... | 20 |
| 表 3 集成模型混淆矩阵表..... | 36 |
| 表 4 集成模型效果评分表..... | 37 |
| 表 5 所有模型准确率比较..... | 37 |

摘要: 初创企业是指刚刚创立且没有足够资金以及资源的各类企业。初创企业面临着各种不确定性，失败率也很高，约 10 个企业中 9 个企业会失败。因此，在初创企业中只有少数企业能够成功，并且其中的极少部分企业能够成为 unicorns。本次作业分别使用 Kaggle 网站的两个初创企业数据集，进行不同的研究：对第一个数据集进行探索式数据分析，总结出成功初创企业的特点，描绘成功初创企业的画像，从而为创业者提供创业建议；对第二个数据集，使用机器学习的方法，建立能够用来预测初创企业成功率的二元分类模型，并进行模型优化与评估，从而为投资者提供投资方向。

关键词: 初创企业；成功率；探索式数据分析；机器学习；二元分类；调参；

针对初创企业成功率的分析与预测

一、选题背景

1. 1. 研究背景

初创企业是指刚刚创立且没有足够资金以及资源的各类企业^[1]。在政府以及学校的支持下，越多的人愿意以创立企业追求他们的梦想^[2]。而且，创业公司正在成为推动我们经济以及科技等多方面发展的一种重要手段。例如，苹果或者最近的美国企业 uber 都为社会、经济以及科技发展具有巨大的影响力。然而，初创企业面临着各种不确定性，以及失败率也相当高^[3]。Hanley 指出，在 10 个企业中约有 9 个企业很大可能会面临失败^[4]。因此，对于初创企业，了解并分析导致初创企业成功与失败的因素至关重要^[5]。加上，若能够以该企业的特征来预测初创企业的成功与否，将会成为预测企业成功或失败的重要手段。

[1] Rebecca Baldridge, and Benjamin Curry. 2021. What is a Startup? *Forbes* (April). <https://www.forbes.com/advisor/investing/what-is-a-startup/>

[2] Kim, E. (2015). Fastest startups to \$1 billion valuation - Business Insider. Retrieved August 21, 2017, from <http://www.businessinsider.com/fastest-startups-to-1-billion-valuation-2015-8/#slack-is-the-fastest-growing-enterprise-software-ever-11111114>

[3] Ries, E. (2011). The Lean Startup. *Working Paper*, 1–28. <http://doi.org/23>

[4] Hanley, J. A., & McNeil, B. J. (1982). The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. *Radiology*, (143). Retrieved from <http://pubs.rsna.org/doi/pdf/10.1148/radiology.143.1.7063747>

[5] Francisco Ramadas. 2017. Predicting Start-up Success with Machine Learning. University of Lisbon. Published in November <https://run.unl.pt/bitstream/10362/33785/1/TGI0132.pdf>

1. 2. 研究内容

本次作业以初创企业的成功为主题，一共分为两个方向：

其一是对初创企业数据集进行探索式数据分析，将用导致初创企业成功或失败的一些特征来分析出哪些因素下初创企业的成功率高，或者在哪些情况下初创企业的失败率提高，并为创业者提供创业建议；

其二是用机器学习的方法，选取初创企业的一些特征来建立模型。在本文的模型建立过程中，将会使用未调超参数的多个模型进行模型比较，最终选取其中模型效果最高的两个模型进行调优参数。调完参后再次进行模型比较，此期间会使用交叉验证（cross validation）方法避免模型的过拟合现象。最后，选取效果最好的两个模型再次建立 Voting Classifier，并给出最终的模型效果。该预测结果会为创业者提供创业指南的同时，也会为投资者提供投资方向。

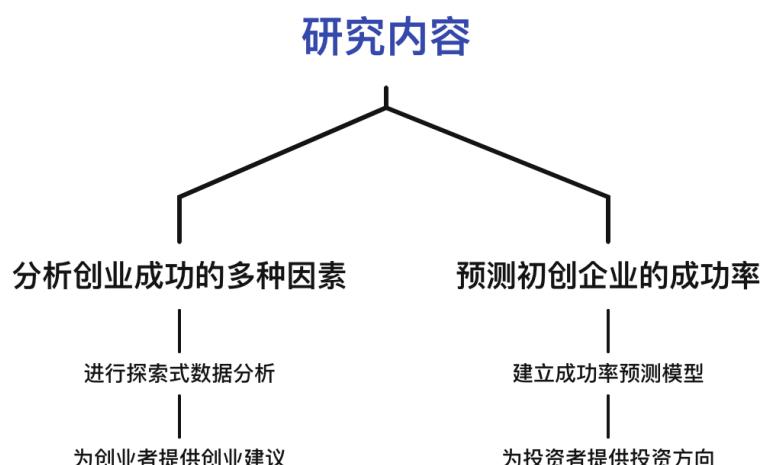


图 1 研究方向思维导图

1. 3. 研究意义

如在研究背景所述，由于初创企业的失败率相当高，并且，由于不同研究使用不同的数据集，也即选取不同特征以机器学习的方法对初创企业成功与否进行预测^[6]，因此，本研究中使用的特征和最终结果会与以往研究不同，这也将会为初创企业开阔新的视野，为初创企业提供更多的余地去思考并改进。

[6] Anna Melnychuk. 2021. STARTUP SUCCESS PREDICTION. EXAMPLE FROM THE US. <https://kse.ua/wp-content/uploads/2021/12/Anna-Melnychuk-1>

二、成功企业画像

2.1. 数据介绍

2.1.1. 数据来源

为了全面展现成功企业不同方面的特质，我们选择了 Kaggle 上提供的 startup-analysis 数据集进行数据分析。该数据集字段非常丰富，样本数量较少，适宜做纯探索式数据分析。

网址：<https://www.kaggle.com/datasets/ajaygorkar/startup-analysis>

2.1.2. 数据检视

startup-analysis 数据集按照公司上市或被收购划分为成功公司、破产为失败公司的标准将公司进行划分，并提供对应公司的时空信息、产业类别、创业者个人素质、创业团队素质及公司福利相关信息。

- 数据集的样本数达到 472，总字段数为 116。

我们的目标字段为企业成功与否：Dependent-Company Status，取值为 1 表示成功，为 0 表示失败

- 将所有特征字段进行分类，可以分成如下八类：

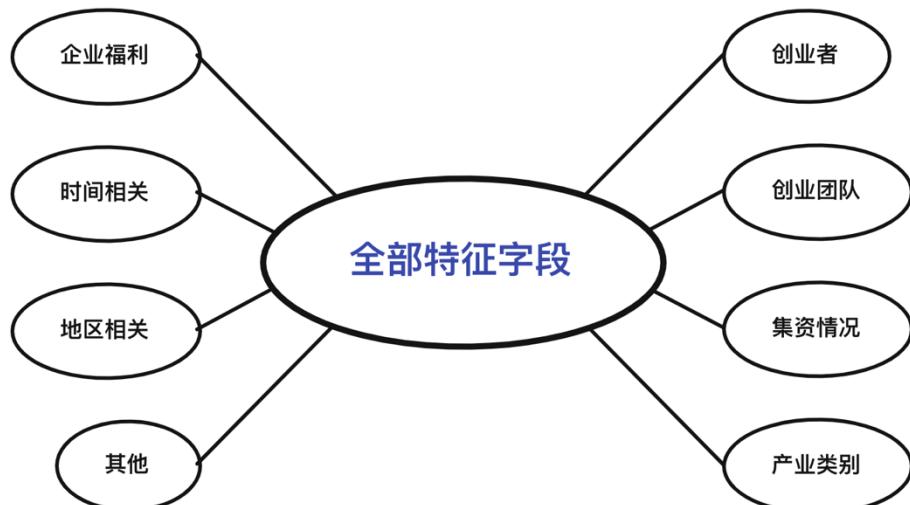


图 2 全部特征字段分类

由于对于一部分字段含义尚不明确或者与分析无关，因此对字段进行筛选。

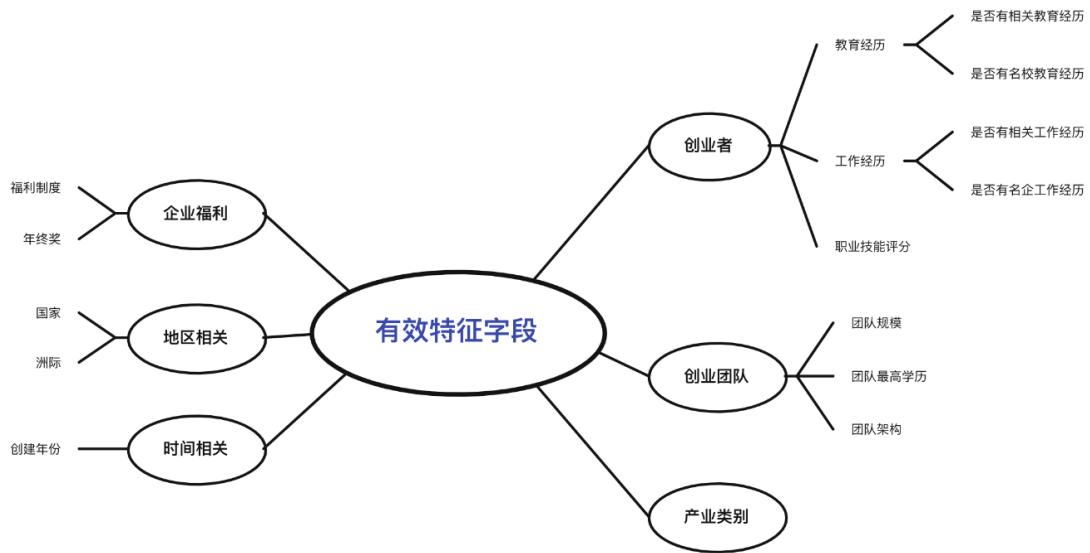


图 3 有效字段筛选

根据有效特征的字段，我们选取以下字段作为成功企业画像的主要指标：

- 企业基础信息：地区分布、创建年份、产业类别

| | | | |
|---------------------|------------------|--------------------|----------------------|
| Industry of company | year of founding | Country of company | Continent of company |
|---------------------|------------------|--------------------|----------------------|

- 创业者画像：是否有名企实习经历、是否有相关工作经历、是否毕业于名校、是否有相关教育经历、职业技能评分

| Worked in top companies | Relevance of experience to venture | Degree from a Tier 1 or Tier 2 university? | Relevance of education to venture | Skills score |
|-------------------------|------------------------------------|--|-----------------------------------|--------------|
|-------------------------|------------------------------------|--|-----------------------------------|--------------|

- 创业公司画像：团队规模、团队最高学历、团队架构、企业薪资结构、企业有无激励奖

| Employee Count | Employees count MoM change | Number of Co-founders | Number of advisors | Team size Senior leadership |
|-------------------------|---|-----------------------|------------------------|-----------------------------|
| Team size all employees | Number of Recognitions for Founders and Co-founders | Highest education | Team Composition score | |

| | |
|---|----------------|
| Employee benefits and salary structures | Company awards |
|---|----------------|

2. 2. 企业基础信息画像

在基础信息画像这一部分中，我们主要关注该数据集的基本信息，包括企业运营情况（成功还是失败）、时空分布、公司的产业类型等。

2. 2. 1. 公司运营情况

本数据集中，关于企业成功的定义为“被其他公司收购或上市”，认为企业失败的条件为“企业宣布破产”。在本数据集中，成功公司共有 305 家，占该数据集中公司数的 64.62%，失败公司共有 167 家，占全部公司数的 35.38%。

考虑到两种类型企业的数量差异，因此在我们后续分析过程中，想要研究单一元素对企业的影响时，选择的比率都是某一类型（成功或失败）的企业，元素的某一类别的占比。

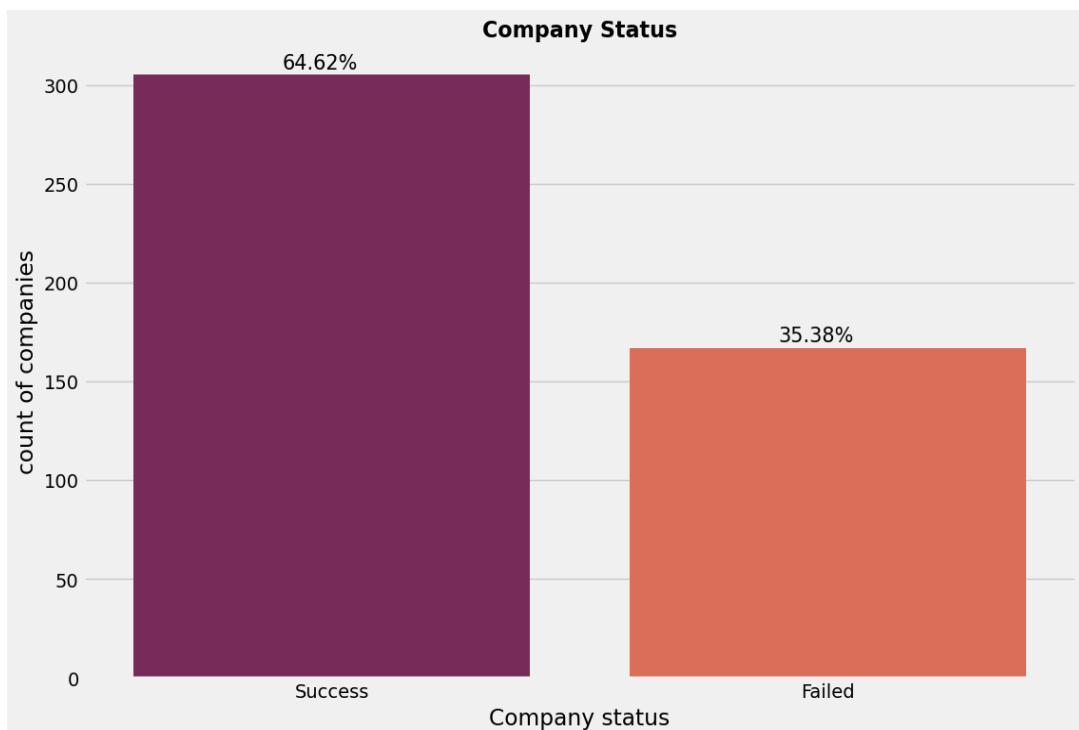


图 4 该数据集中成功与失败企业占比情况

2. 2. 2. 公司的时空分布

通过直方图可以发现，该数据集中的公司成立年份主要集中在 2007 年及以后，占该数据集总数的 75.1%；而从地域上来看，此数据集调查的公司主要分布在美国和英国，占所有公司总数的 71.6%。

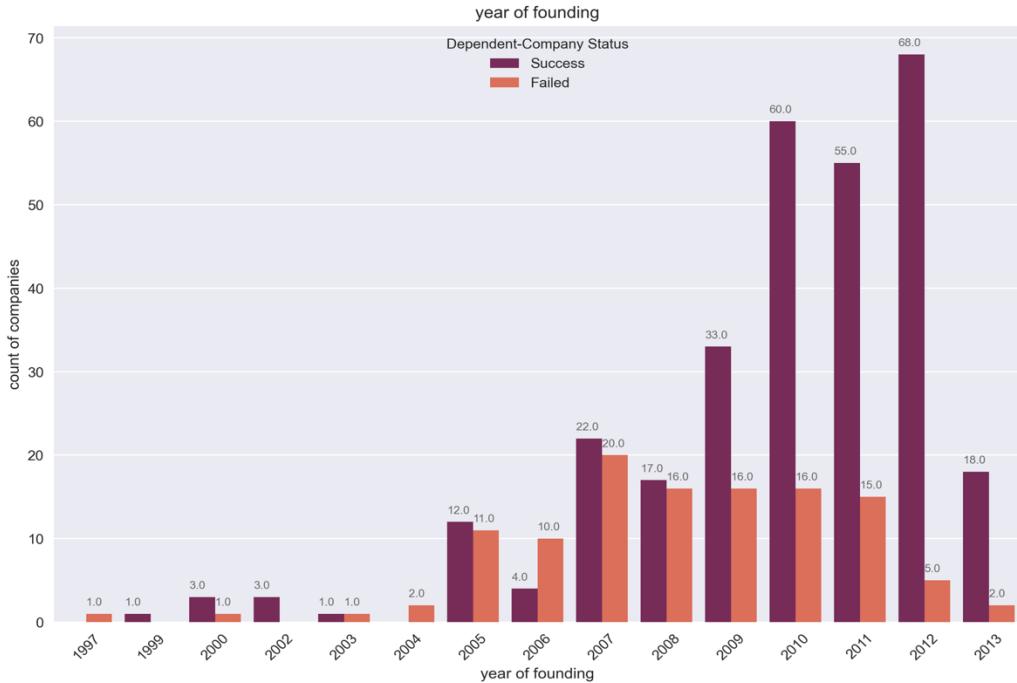


图 5 数据集中不同公司成立年份分布

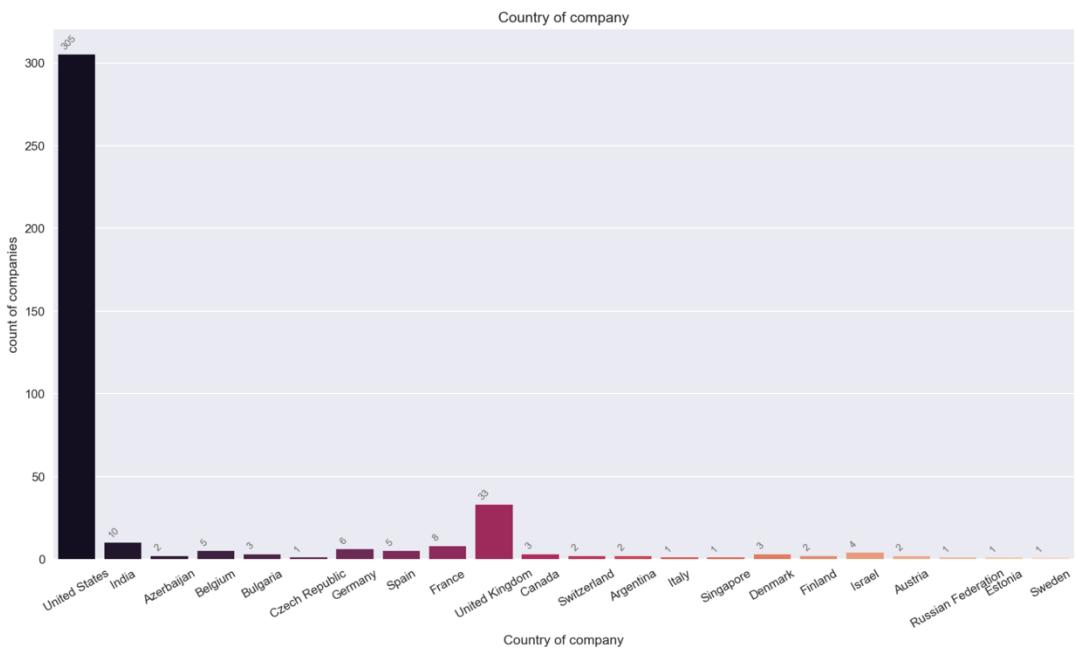


图 6 数据集中不同公司的地区分布

2.2.3. 公司的产业类型分布

如图 7、图 8 所示，从整体来看，创业的产业类型呈现了偏向电子、互联网产业的趋势。分析(analytics)、移动通信(mobile)、电子商务(e-commerce)似乎已经占据创业市场的主流，不过尽管电子、互联网产业发展迅猛，但是同样值得注意的是，销售行业(marketing)仍然占据一定的市场。说明当时时代背景下，创业市场呈现出传统行业逐渐衰落但占据一定地位，而互联网技术、电子产业发展迅猛的迹象。

而对于成功企业来说，占据其前三位的产业类型分别为：分析（占比 68.8%）、电子商务（占比 23.0%）市场营销（占比 18.7%）；而对于失败企业来说，占据其前三位的企业类别分别为：分析（占比：26.5%）、移动通信（占比：15.1%）、电子商务（占比：14.1%）。

考虑到因为在 2010 年左右，移动通信尚且处于初步发展阶段，因此移动通信在当时的风脸较大，可能是在失败企业中占比比较高的原因。

另外，我们发现，在成功企业与失败企业中，分析与电子商务都占据了比较大的比重，为了探究两类产业在成功与失败企业之间是否存在显著性差异，我们又进行了卡方独立性检验，卡方值为 3.93， $p = 0.04$ ，说明电子商务以及分析行业在失败企业和成功企业之间有显著性的差异，成功企业中分析行业、电子商务行业占比更高。

尽管我们选取的占比是组内占比（例：成功企业中分析产业占比、失败企业中分析行业占比），考虑到成功企业与失败企业的数量差距，采用分样本取样的方式仍具有一定说服力，这也在一定程度上为我们提供了一些创业建议，即可以考虑选择一些发展前景较好且刚刚起步、存在一定规模的产业进行创业。

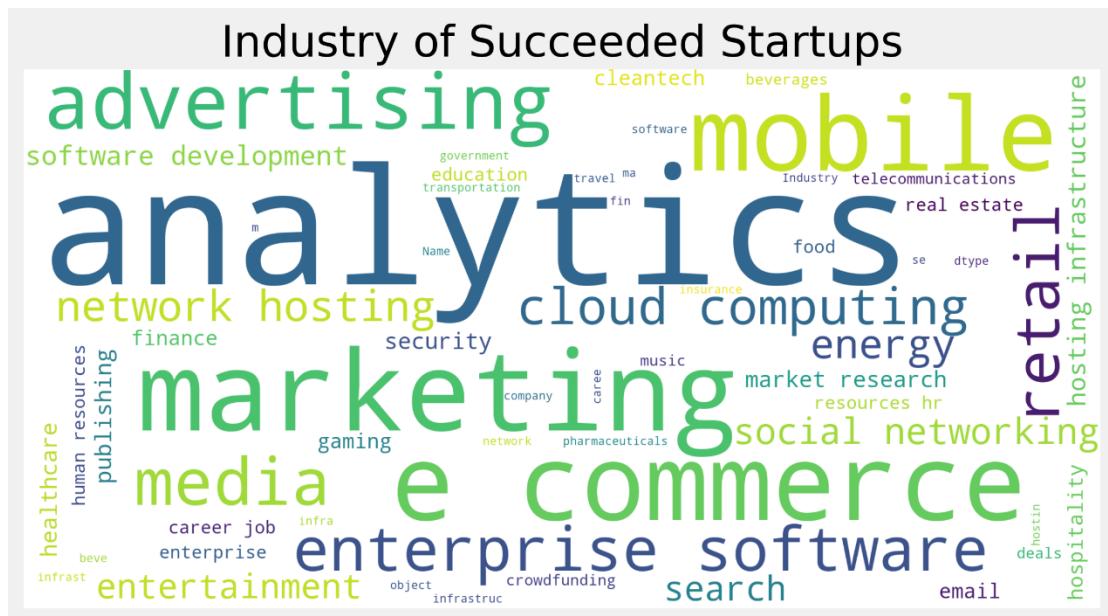


图 7 成功企业产业类别词云图

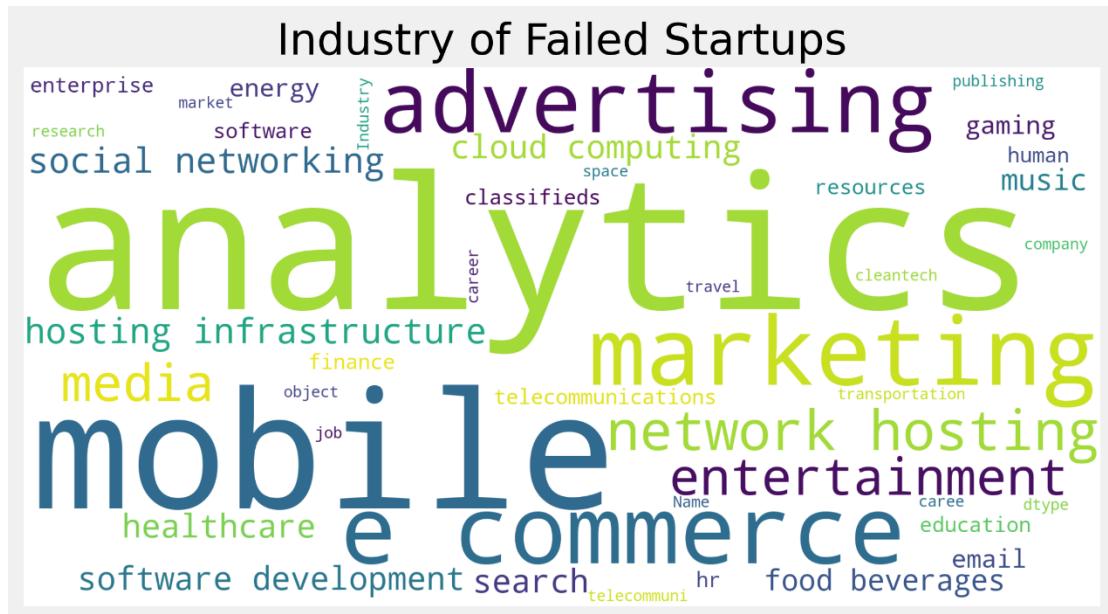


图 8 失败企业产业类别词云图

2.3. 创业者画像

在这一部分，我们重点关注成功与失败企业中创业者的个人特质之间的差异，包含创业者的相关经历、创业者的受教育程度等。

2.3.1 创业者的工作经历

从图 7 我们可以看出，成功企业的创业者具 top 公司工作经历或具有相关工作经历的占比更高。

对于该两种进行两次卡方独立性检验，结果表明：对于 top 公司工作经历，卡方值为 1.81， $p=0.17$ ，说明是否有 top 公司的工作经历在两类公司中并不独立；对于相关工作经历，卡方值为 3.36， $p=0.06$ ，属于边缘显著，有理由认为具有相关工作经验可能会在成功与失败企业之间不独立，成功公司的创业者具有相关工作经历占比更高。

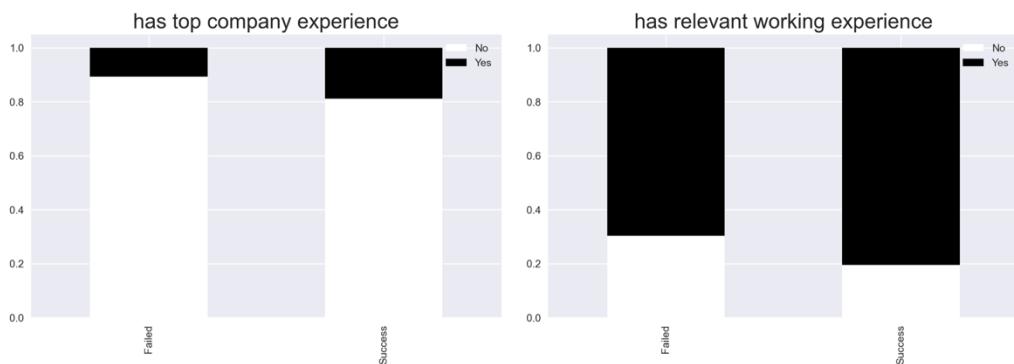


图 9 创业者工作经历堆积图

2.3.2 创业者的受教育程度

针对创业者的毕业学校类别，由图 8 可以看出，对于成功企业而言，其创业者毕业于一类大学(Tier_1)和二类大学(Tier_2)以及毕业生群体毕业于一类大学和二类大学占比更高，说明创业者的受教育程度对创业成功影响较大。

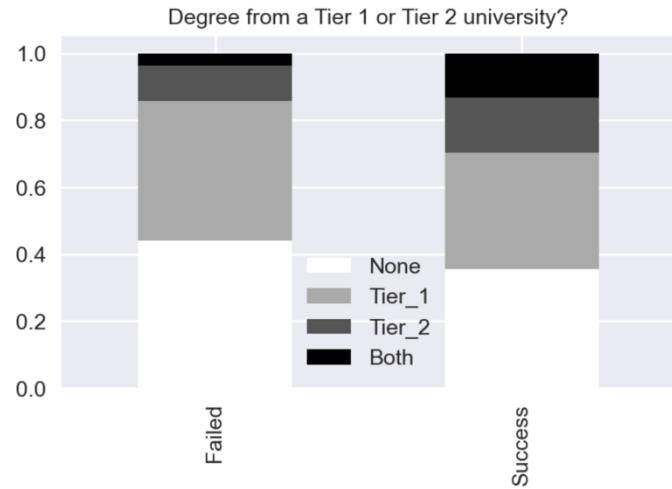


图 10 创业者毕业学校堆积图

针对四类在不同类别的企业占比情况，进行了卡方独立性检验，检验结果为，卡方值为 9.01， $p=0.03$ ，说明该四类在两类企业的分布有显著性差异，为了探究哪些水平在两类公司有显著性差异，又进行了事后检验(post-hoc)，检验结果采用 FDR 方法进行校正。校正结果见表 1。

| | 比较对象 | p 值 | 矫正后 p 值 | 显著性 | 拒绝情况 |
|---|------------------|----------|----------|-----|-------|
| 0 | (None, Tier_1) | 0.970372 | 0.970372 | ns | False |
| 1 | (None, Tier_2) | 0.169113 | 0.292212 | ns | False |
| 2 | (None, Both) | 0.015747 | 0.056036 | * | False |
| 3 | (Tier_1, Tier_2) | 0.194808 | 0.292212 | ns | False |
| 4 | (Tier_1, Both) | 0.018679 | 0.056036 | * | False |
| 5 | (Tier_2, Both) | 0.316998 | 0.380398 | ns | False |

表 1 不同类别学校毕业的时候比较列表

尽管两两事后比较结果并不显著，但是据原 p 值可以得到，None 与 Both 组，Tier_1 与 Both 组均有显著性差异。结合图 8 我们可以发现，成功的公司创业者无大学文凭或毕业于 Tier_1 大学占比显著较低，但是 Both 的比率显著更高。

针对此类现象，我们认为：

- (1) 首先教育程度对创业较为重要，没受到高等教育创业失败率可能较高；
- (2) 教育的多样性较为重要，Tier_1 学校可能培养的是学生的思维以及学术能力，但是 Tier_2 学校的学生社会互动性、对社会了解程度可能更高，因此在进行创业时可以形成一个团队，涵盖较全的教育背景。

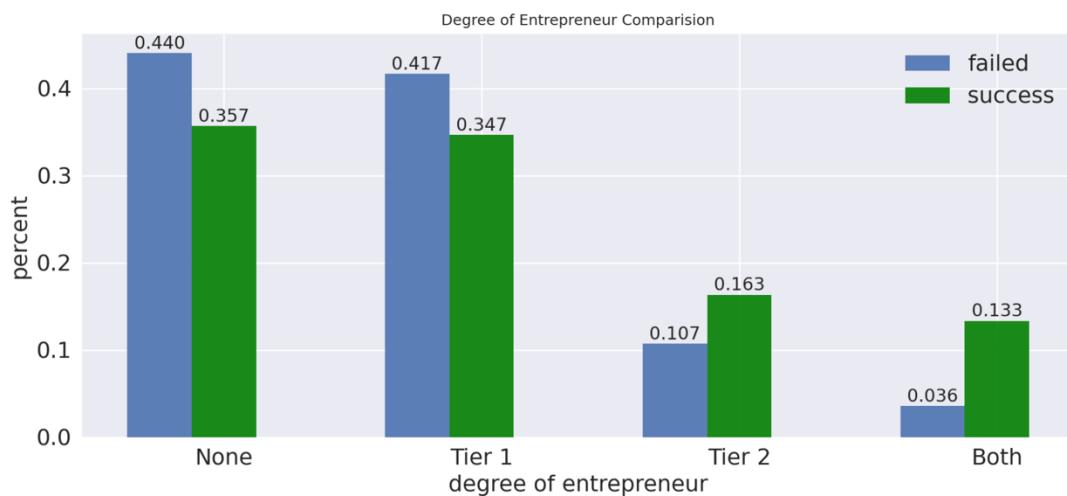


图 11 两类公司中创业者毕业学校堆积图

此外，针对创业者的受教育时长，笔者进行了独立样本 t 检验，结果显示：莱文检验结果为 3.06， $p = 0.08$ ，说明数据满足正态性；对于 t 检验的结果， $M = 0.68$ ， $p = 0.49$ ，说明成功企业与失败企业的创业者在教育时长上没有显著的差别。这一结果表明，受教育时长对创业成功与否影响并不显著。

2.3.3 创业者的技能得分情况

最后，针对创业者的技能评分(包括业务能力、技术水平等)，由图 12 和图 13 同样可以看出，成功企业创业者的技能得分显著更高。

对两类公司的创业者进行独立样本 t 检验，结果显示：莱文检验结果为 2.897， $p = 0.09$ ，说明数据满足正态性；对于 t 检验的结果， $M = -0.92$ ， $p = 0.35$ 。尽管统计学上两者并不存在显著的差异，但是在数字上已经呈现出此类趋势。说明想要创业成功，需要提升创业者本身的个人能力，即“打铁还需自身硬”。



图 12 成功企业与失败企业创业者技能得分的密度分布图

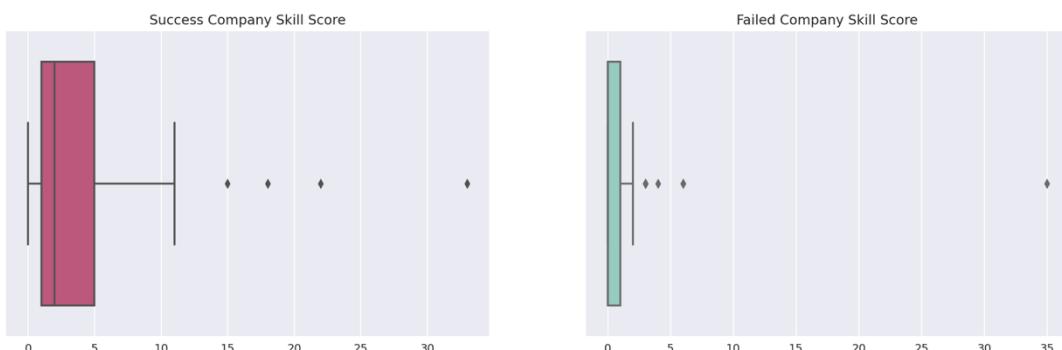


图 13 成功企业与失败企业创业者技能得分的箱线图

2.4. 创业公司画像

在这一部分，我们重点关注成功与失败企业中创业公司整体间的差异，包含创业团队的规模、团队组成、创业团队的最高学历、创业公司的福利机制等。

2.4.1. 创业公司的规模

对于公司的规模类数据，由图 14 可以看出，成功公司的人数规模普遍更高，包括顾问数、员工数、联合创始人个数。此外，也可以看出，成功公司人数的变化情况也在逐步增加，员工数目月增加率大于 0.

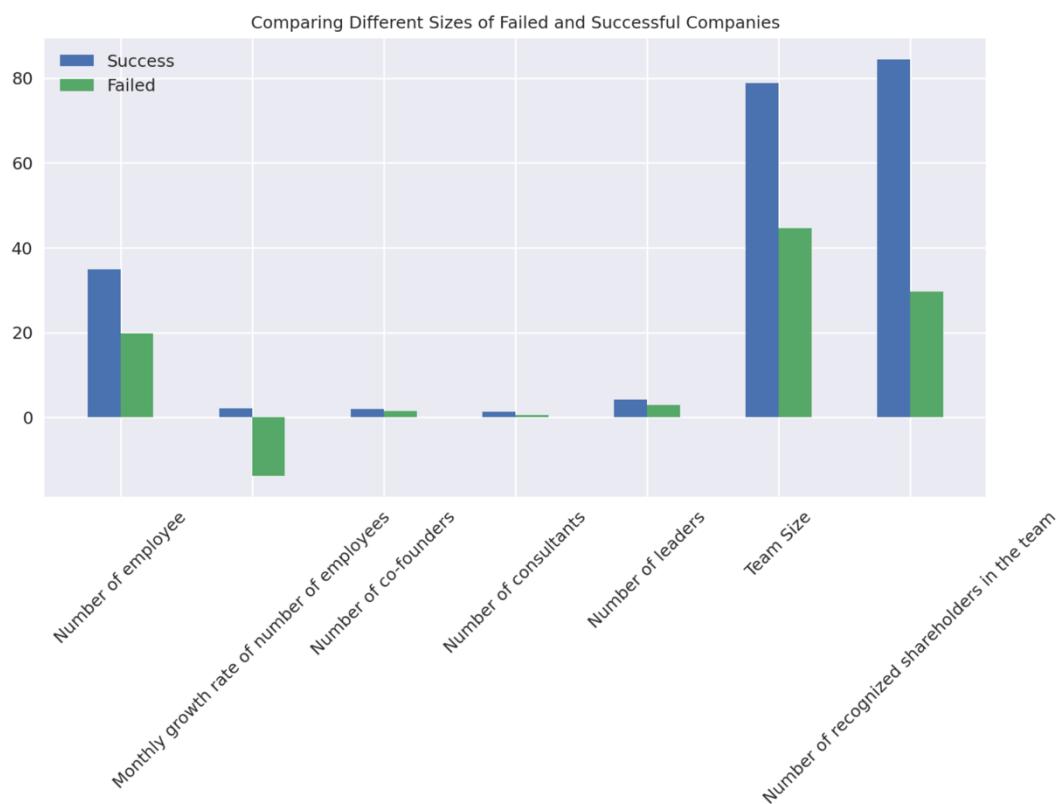


图 14 成功企业与失败企业的公司规模

2. 4. 2. 创业公司的团队组成

结合 2. 4. 1 的分析以及图 15 可以看出，虽然成功公司人数更多，但是并没有影响公司的架构。从图 15 中我们可以看出对于成功公司而言，公司结构评分更高。其中，结构评分为高的占 24. 9%，中等的占 34. 6%，而评分为低的仅有 40. 5%；而对于失败公司，结构评分占高中低的比率分别为 8. 0%、19. 5%、72. 4%。

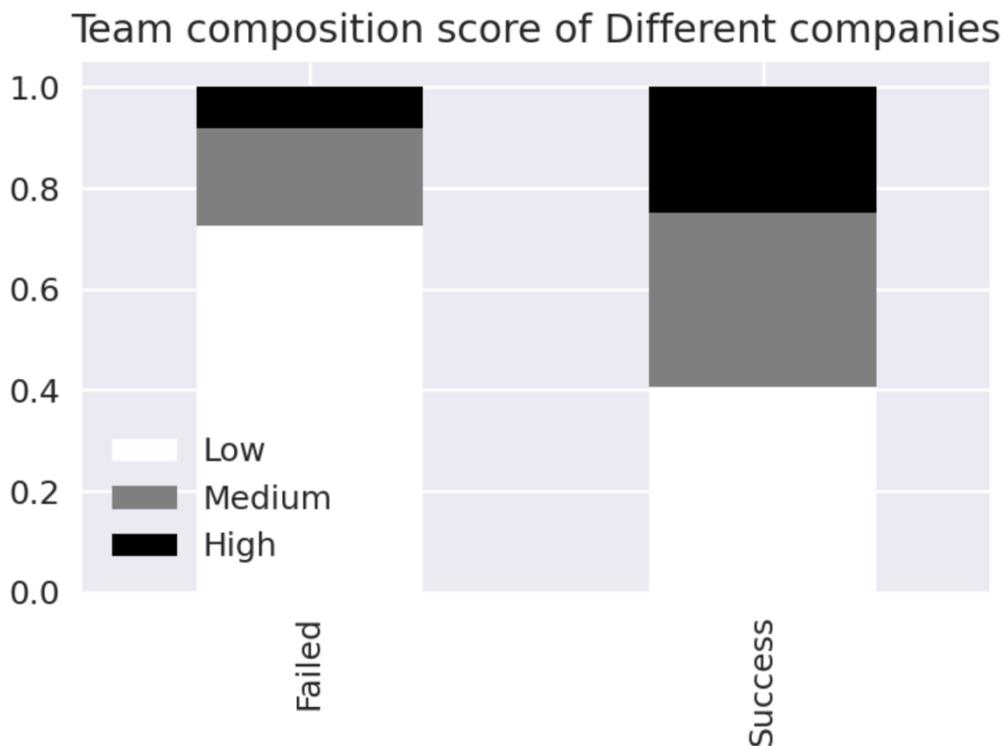


图 15 成功公司与失败公司结构评分情况

结合 2. 4. 1、2. 4. 2 我们可以认为，对于成功公司，其一个经典特征为公司人数更多且结构更加合理。因此对于创业公司来说，一方面需要提高人员数量，提高生产力；另一方面需要优化产业结构，提高产业效率。

2. 4. 3 创业公司的最高学历

对于企业的创业团队，笔者还统计了企业中最高学历分布情况。由图 16、17 可以看出，失败企业与成功企业学士比例差距不大，均在 45% 左右浮动；而失败企业中硕士占比更高，为 52.56%，成功企业中博士占比更高，为 10.65%。

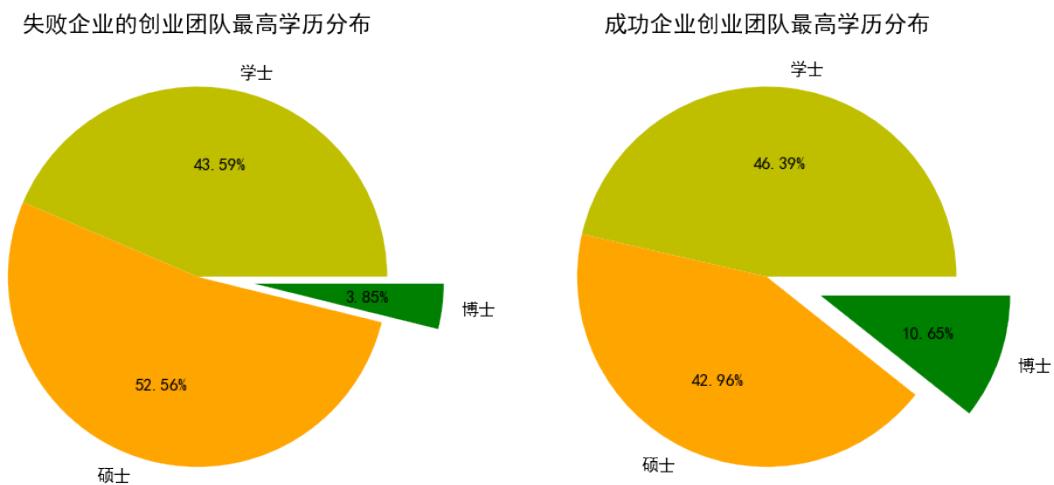


图 16 成功公司与失败公司创业团队中最高学历分布饼图

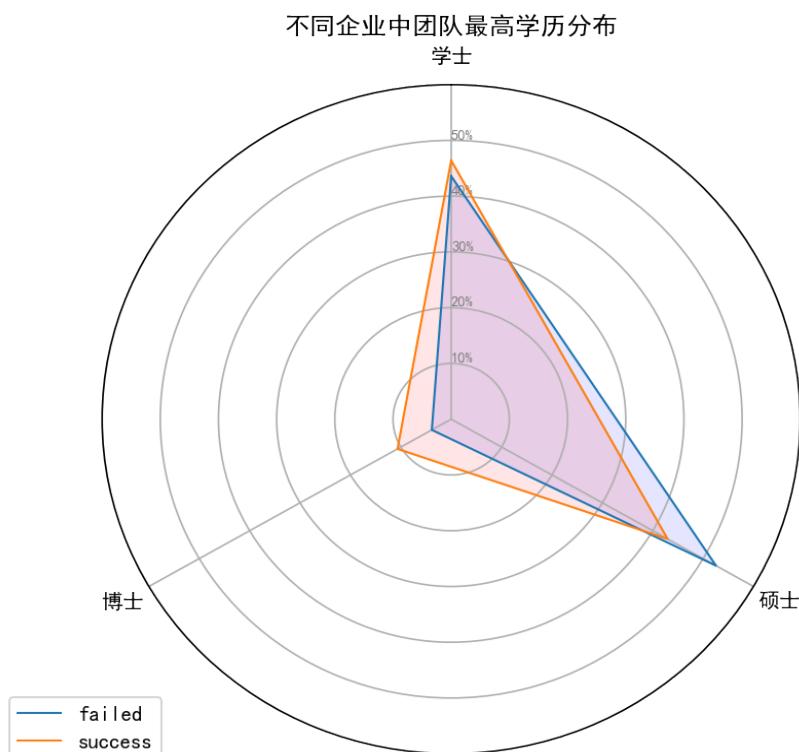


图 17 成功公司与失败公司创业团队中最高学历分布雷达图

从人才培养的角度出发，将三种类型的人才划分成技术型人才（包含学士和硕士）、科研型人才（包含博士），重新绘制饼状图并进行卡方独立性检验，结果表明：卡方值为 2.64， p 值为 0.10， α 取 0.1 可以认为边缘显著，说明成功企业的科研人才可能比失败企业中统计学意义上更多。

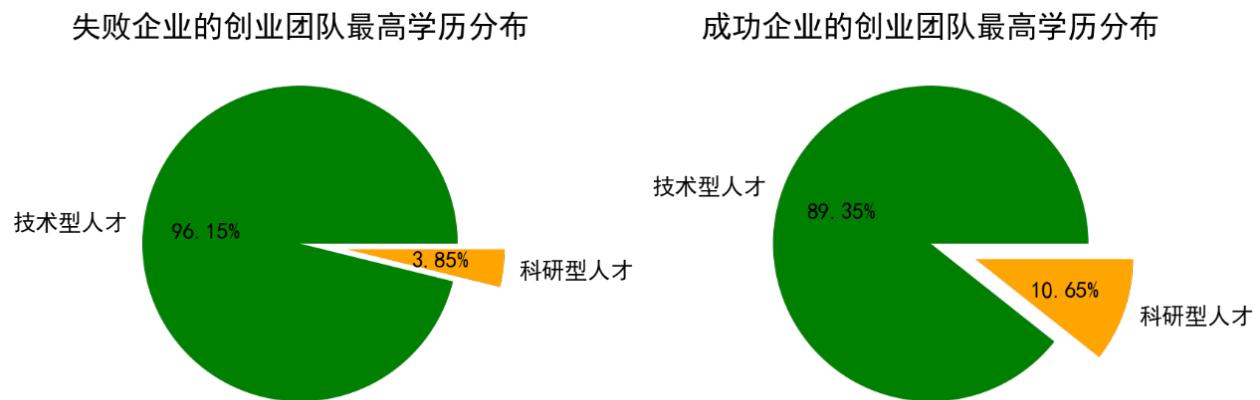


图 18 成功公司与失败公司创业团队中人才类型分布饼图

针对此，联系到词云分布，可以认为：高新技术的崛起同样意味着技术迭代的加快，对技术商品的创新性要求进一步提高，因此，适度提高企业中科研人才的比率，提高企业创新能力具有一定的帮助。

2. 4. 4. 创业公司的福利机制

对成功与失败公司的薪水结构同样进行分析，结果发现，对于成功企业，其评价为‘Very Good’以及‘Good’的比率更高；而对于失败企业，评价为‘Bad’的比率更高。

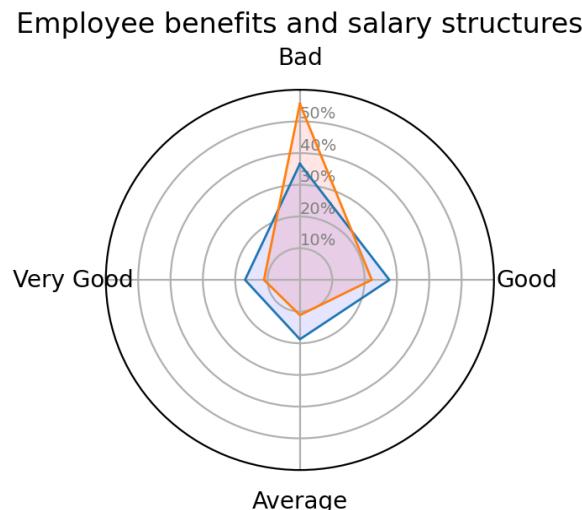


图 19 成功公司与失败公司创业团队中薪资结构评价

同样对于两类企业的奖励机制进行了比较，同样发现，对于成功企业，其具有奖励机制占比更高，而该数字在失败企业中仅为 8.0%。

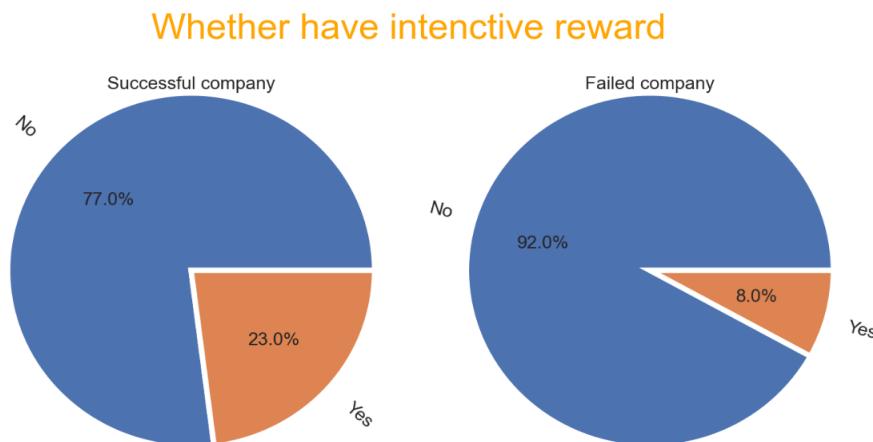


图 20 成功公司与失败公司创业团队中的奖励机制

综合薪资结构与奖励机制可以发现，成功企业的薪资结构更高，激励机制更多。所以想要保障企业的正常运行，一个合理的建议是，切实保障企业的福利机制，从而调动员工的积极性。

2.5. 小结

2.5.1. 结果总结

本次EDA，我们从企业创业者、创业公司两个方面绘制成功初创企业画像。

- 在企业创业者方面：

- (1) 相关经历：成功企业的创业者往往更具备相关经验。相较于失败公司，成功企业创业者具有名企工作经历、相关工作经历占比更高；
- (2) 受教育程度：相较于失败企业，成功企业的创业者受教育质量与丰富度更高，但是受教育时长并不存在显著性差异；
- (3) 技能得分：成功企业的创业者得分相对更高，其个人更具有组织协调与业务能力。

- 在创业公司方面：

- (1) 团队规模：成功企业的规模更大，员工人数、顾问人数更多，且企业人数呈现上升趋势。
- (2) 创新能力：成功企业的科研型人才人数占比更高，创新能力更高
- (3) 团队组成：成功企业的团队组成得分更高，说明成功企业结构更加合理。
- (4) 福利机制：成功企业的薪资水平更合理，具有激励机制的企业占比更高。

2.5.2. 相关建议

结合5.1部分结果总结，可对创业者或预创业者提供如下建议：

- (1) 首先重视个人能力，可通过接受更高质量的教育、积累相关经验进一步提高自身的能力；
- (2) 注重团队力量：在组建创业团队时可以联络不同背景的伙伴；组建公司时也可以吸引优秀创新型人才；
- (3) 完善公司结构：组建公司时可以适度提高公司规模，提高员工数目，同时完善公司结构；同时完善公司福利机制，保障员工物质需求。
- (4) 筛选创业方向：需要结合当时市场风向，如果是传统产业可能需要进行适度的改革，如果是新兴产业需要评估风险与该产业的未来发展。

三、企业成功率预测模型

3.1. 数据介绍

3.1.1. 数据来源

为了构建效果较好的分类模型，我们选择了 Kaggle 上提供的 startup-success-prediction 数据集进行数据分析。该数据集样本数量较多，字段情况比较符合训练规范，适宜做机器学习分类模型的构建。

网址：<https://www.kaggle.com/datasets/manishkc06/startup-success-prediction>

3.1.2. 数据检视

startup-success-prediction 数据集按照公司被其他机构收购为成功公司、公司关闭为失败公司的标准将公司进行划分，并提供对应公司的时空分布信息、投资信息和产业类别信息等特征信息。

- 数据集的样本数达到 923，总字段数为 49。

我们的目标字段为企业成功与否：labels，取值为 1 表示成功，为 0 表示失败。

3.1.2.1 特征字段详情

| 字段名称 | 数据类型 | 字段含义 | 具体类型 |
|------------------|--------------|----------|-------------|
| Unnamed: 0 | quantitative | 无明确含义 | |
| state_code | categorical | 企业所在州的编码 | multiple |
| latitude | quantitative | 企业位置纬度 | |
| longitude | quantitative | 企业位置经度 | |
| zip_code | categorical | 企业所在地邮编 | multiple |
| id | categorical | 企业 id | multiple |
| city | categorical | 企业所在城市 | multiple |
| Unnamed: 6 | categorical | 无明确含义 | |
| name | categorical | 企业名称 | multiple |
| labels | categorical | 企业成功与否 | binary |
| founded_at | categorical | 建立时间 | time series |
| closed_at | categorical | 关闭时间 | time series |
| first_funding_at | categorical | 首轮集资时间 | time series |
| last_funding_at | categorical | 末轮集资时间 | time series |

| | | | |
|--------------------------|--------------|---------------------|----------|
| age_first_funding_year | quantitative | 首轮集资时企业的年龄 | |
| age_last_funding_year | quantitative | 末轮集资时企业的年龄 | |
| age_first_milestone_year | quantitative | 企业达到首个里程碑时企业的年龄 | |
| age_last_milestone_year | quantitative | 企业达到最后一个里程碑目标时企业的年龄 | |
| relationships | quantitative | 企业与外界人士建立联系的数量 | |
| funding_rounds | quantitative | 集资轮数 | |
| funding_total_usd | quantitative | 总共集资的金额(美元) | |
| milestones | quantitative | 企业自己设置的里程碑数量 | |
| state_code.1 | categorical | 同 state_code | multiple |
| is_CA | categorical | 是否位于加州 | binary |
| is_NY | categorical | 是否位于纽约 | binary |
| is_MA | categorical | 是否位于麻省 | binary |
| is_TX | categorical | 是否位于德州 | binary |
| is_otherstate | categorical | 是否位于其他州 | binary |
| category_code | categorical | 企业产业类别的名称 | multiple |
| is_software | categorical | 是否为软件开发企业 | binary |
| is_web | categorical | 是否为互联网企业 | binary |
| is_mobile | categorical | 是否为移动通信企业 | binary |
| is_enterprise | categorical | 是否为未注册中小型企业 | binary |
| is_advertising | categorical | 是否为广宣企业 | binary |
| is_gamesvideo | categorical | 是否为游戏开发企业 | binary |
| is_ecommerce | categorical | 是否为电商企业 | binary |
| is_biotech | categorical | 是否为生物科技企业 | binary |
| is_consulting | categorical | 是否为咨询企业 | binary |
| is_othercategory | categorical | 是否为其他类别企业 | binary |
| object_id | categorical | 同 id | multiple |
| has_VC | categorical | 是否有风险投资家投资 | binary |
| has_angel | categorical | 是否有天使投资者投资 | binary |
| has_roundA | categorical | 是否有 A 轮集资 | binary |
| has_roundB | categorical | 是否有 B 轮集资 | binary |
| has_roundC | categorical | 是否有 C 轮集资 | binary |
| has_roundD | categorical | 是否有 D 轮集资 | binary |
| avg_participants | quantitative | 平均每轮参与投资的人数 | binary |
| is_top500 | categorical | 是否为 500 强企业 | binary |
| status | categorical | 企业是否被接受(与成功与否一一对应) | binary |

表 2 机器学习数据集特征字段详情列表

3.1.2.2 缺失值情况

经检查，缺失值都以 np.nan 的形式存在。

使用 missingno 的 matrix 方法对全部字段的缺失值进行检查：

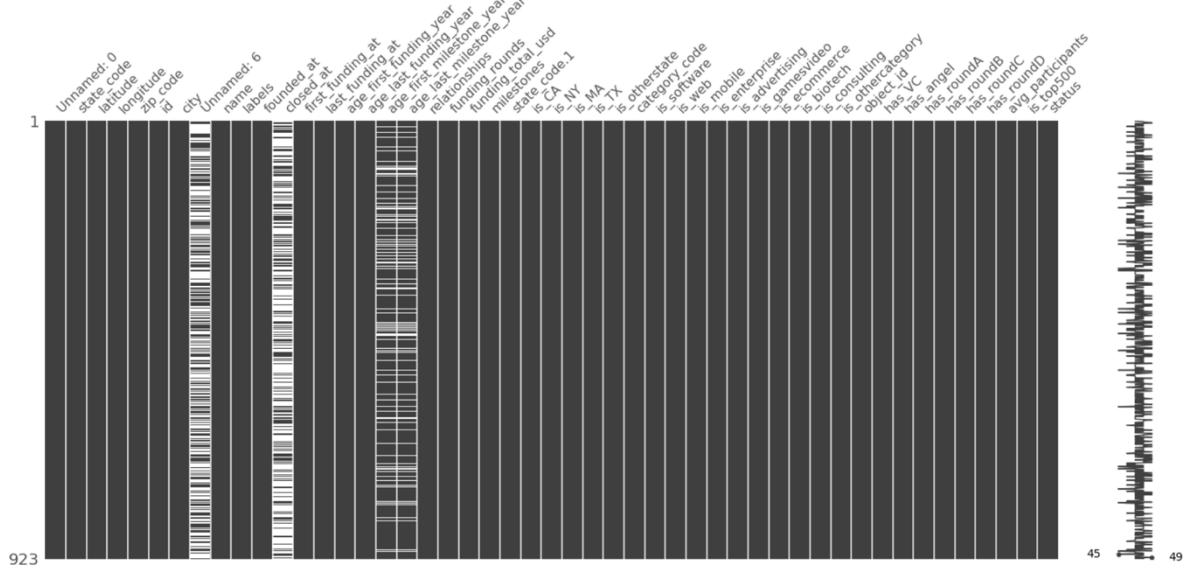


图 21 机器学习数据集缺失值情况分布图

缺失值集中分布于四个字段： `Unnamed: 6` , `closed_at` ,
`age_first_milestone_year` , `age_last_milestone_year` . 其中， `closed_at` 字段缺失值
主要是因为选取的企业样本中有一大部分是没有关闭的企业，所以不存在
`closed_at` 情况；而 `age_first_milestone_year` 和 `age_last_milestone_year` 的缺失值
存在于 `milestones=0` 的情况下，即企业没有设置过里程碑，也就不存在初里程
碑和末里程碑的问题。

3. 2. 探索式数据分析

本部分我们对该数据集中的 10 个重要字段进行单变量分布检视，探索此数据集中样本的分布情况。

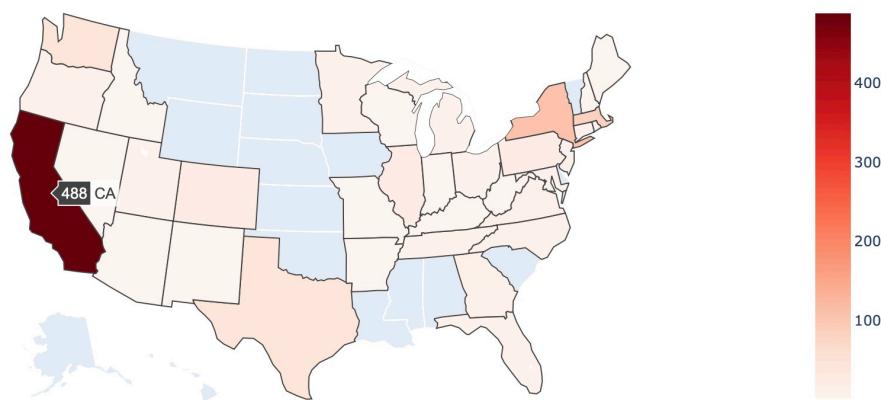
3. 2. 1. 企业的空间分布

本数据集中所有企业均位于 USA，取样的空间分布用其所属州的编码 state_code 表示。

下面对其在美国各州的分布做 Choropleth Map（等值域图）可视化：

等值域图是一种统计专用的地图形式，类似热力图，使用颜色的强度来对应空间查点单位内地理特征的汇总，如人口密度，可以直观地呈现一个变量在整个地理区域的变化情况。

我们利用 python 库 plotly 的 graph_objects 子库，绘制企业分布的等值域图



(将光标放到对应州，便可显示对应州的企业总数)

图 22 机器学习数据集企业地理位置分布等值域图

可见，本数据集中的企业主要集中分布于加州，然后纽约和麻省也有较多分布。东海岸和西海岸各州均有取样，而中部地区几乎没有取样。从数据集采样的地理位置，也可以侧面反映出美国各大州经济的发达程度和创业条件：加州经济发达、高新技术和人才分布集中，且环境条件适宜，是新生企业最为理想的温床；沿海地区经济普遍比内陆发达，也是更加有利于初创企业的建立和发展。

3.2.2. 企业的时间分布

取样的时间分布用 founded_at 字段表示企业创建的年份。

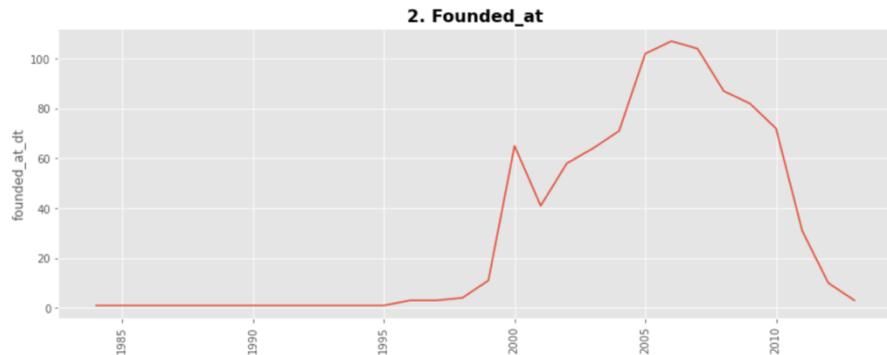


图 23 机器学习数据集企业时间分布折线图

取样时间跨度为 1984-2013，集中分布于 2000 年之后，在 2005-2010 年间达到顶峰。这也侧面反映了步入 21 世纪后，初创企业的增长十分迅速，如同雨后春笋一般，研究初创企业的成功率问题在当下有着重要的现实意义。

3.2.3. 企业的产业类别分布

category_code 表示取样企业的产业类别，下面做了产业类别分布可视化：

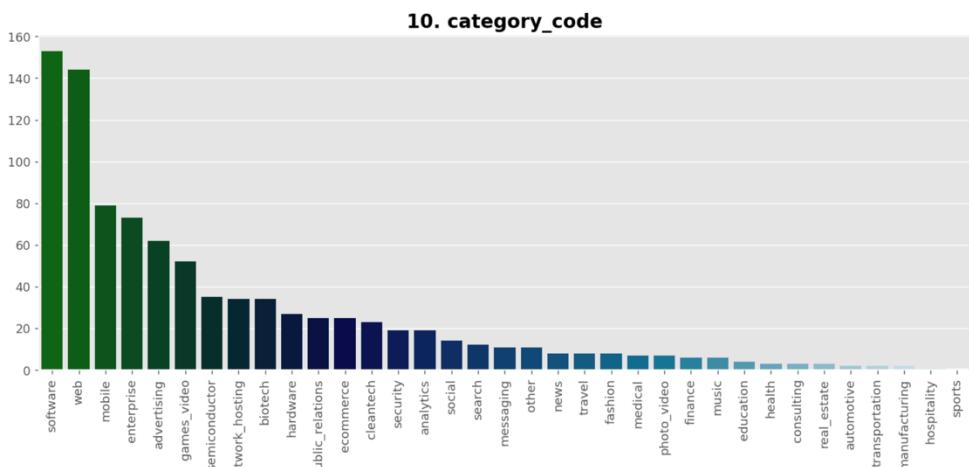


图 24 机器学习数据集企业产业类型分布

可以看出，取样企业中，大多数为软件开发和互联网企业。结合 1.3.2 这也侧面说明了 21 世纪后 IT 行业的崛起，诞生了许多 IT 相关的初创企业。

3.2.4. 企业的投资人情况

投资人情况包括是否有投资人（两类）和参与集资的平均人数三个字段。

- **has_VC & has_angel**

VC 投资家（Venture Capitalist），即风险投资家； angel 投资家（Angel Investor），即天使投资家。VC 投资家和 angel 投资家都是通过为企业提供资金支持并取得企业股份、以期得到回报的客户，但是二者在投资对象和投资金额方面有较大的差异。

在投资对象上，VC 投资家倾向于对建立已有一段时间、有一定发展经验的初创企业进行投资，以降低投资风险，获得更高的报酬；而 angel 投资家则多有对仍处于早期阶段的初创企业进行投资，即使这些企业还没有来得及展现出自己的实力。

在投资金额上，VC 投资家为职业投资人，其投资资金来源于专门的投资公司，所以投资金额数目大；而 angel 投资家多为民间投资人，使用的是私人的资金，因此投资金额也较小。

下面对数据集中企业是否有这两类投资人的占比进行可视化：

（其中，1 表示有该类投资人，0 表示没有该类投资人）

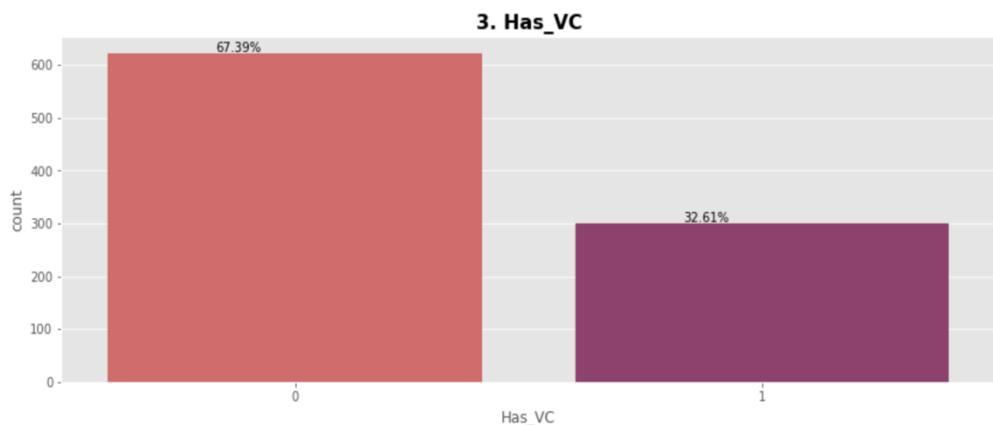


图 25 企业是否有风险投资人柱状图

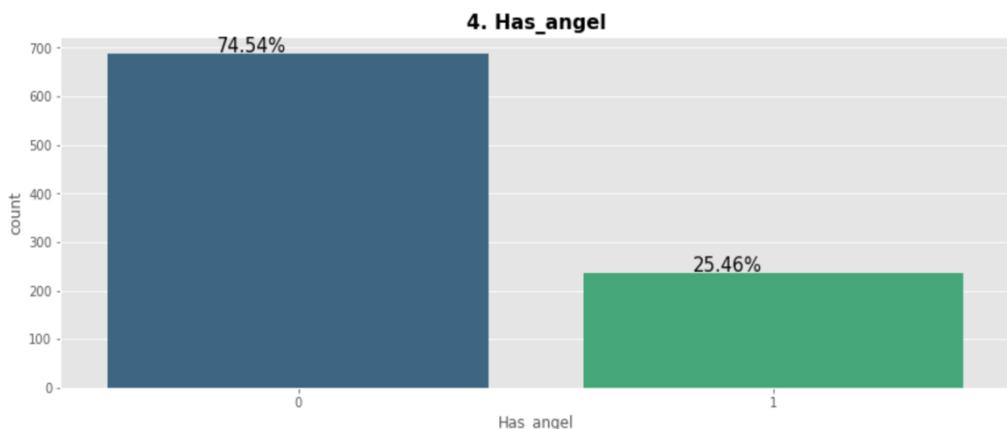


图 26 企业是否有天使投资人柱状图

可以看出，大部分企业没有（或仍未有）投资人。不过，拥有 angel 投资人企业比拥有 VC 投资人的企业略多。

- **avg_participants**

avg_participants 表示平均每轮参与集资的投资者数量。比如，如果一共有 5 轮集资，第一轮集资 1 人参与，第二轮集资 3 人参与，其他轮集资都无人参与，则 $\text{avg_participants} = (1+3)/5=0.8$ 。

下面对该字段做密度分布图

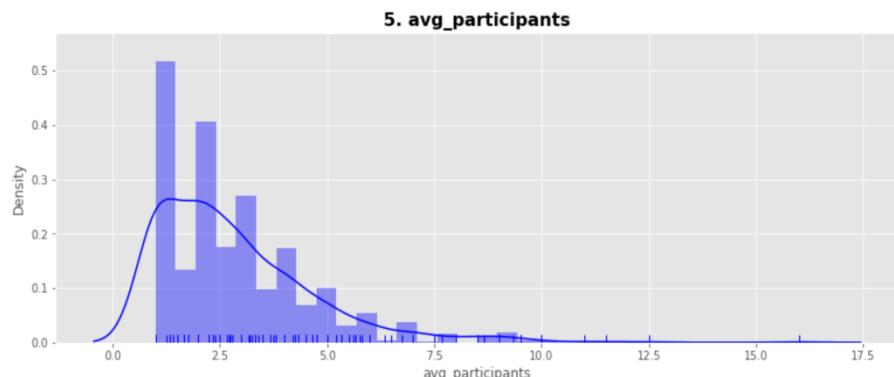


图 27 平均每轮参与集资的投资者数量密度分布图

可以看出，平均每轮参与集资的投资者数量集中分布于 1-2.5 之间，越大越少。

3.2.5. 企业的集资情况

集资情况包括集资总额和融资情况一共三个字段。

- **funding_total_usd**

`funding_total_usd` 表示企业总共集资的金额（以美元为单位）。下面对该字段做箱线图（取以 10 为底的自然对数，横坐标表示金额的数量级）：

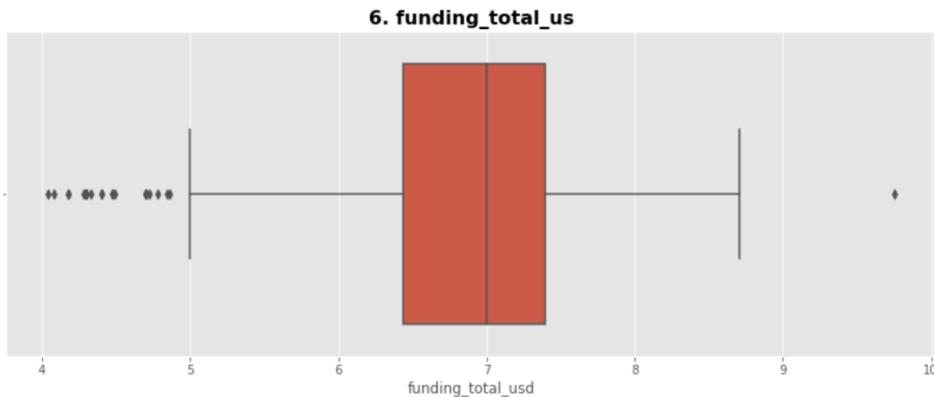


图 28 平均每轮参与集资的投资者数量密度分布图

可以看出，大部分企业的集资金额集中分布于千万美元的量级。其中不乏有过高和过低的离群值：有集资接近百亿级别的企业，可能是发展潜力巨大的高科技，之后发展成为行业独角兽；有集资万元级别的企业，可能是规模较小的非技术型企业。

- **funding_rounds & has_round(A/B/C/D)**

企业在上市之前会经历许多的融资阶段，这些就叫做融资轮次。项目融资属于轮次大致可以分为：种子轮、天使轮、A 轮、B 轮、C 轮、D 轮等。在 A 轮融资阶段，创业公司的产品已经基本成熟，产品上线或者服务已经正常运作一段时间，并有完整详细的商业及盈利模式。度过 A 轮之后，B 轮融资就相对容易。创业公司经过一轮烧钱后，获得了较大发展。甚至一些公司开始盈利，盈利模式趋于完善，可能需要推出新业务、拓展新领域，所以 B 轮资金来源一般是大多是上一轮的风险投资机构跟投、新的风投机构加入、私募股权投资机构加入。到达 C 轮融资的时候，公司已经非常成熟了，除了拓展新业务，就要开始准备上市了，资金来源主要是私募股权投资，有些之前的 VC 也会选

择跟投。更多轮融资的公司大部分是其本身的业务所决定的，有的项目需要大量的烧钱，过早上市不符合这类公司的发展战略，所以需要进行更多轮融资。

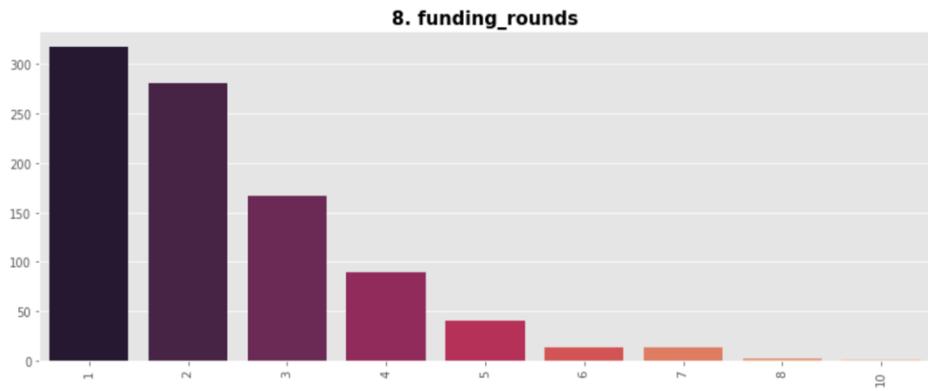


图 29 企业集资轮数分布图

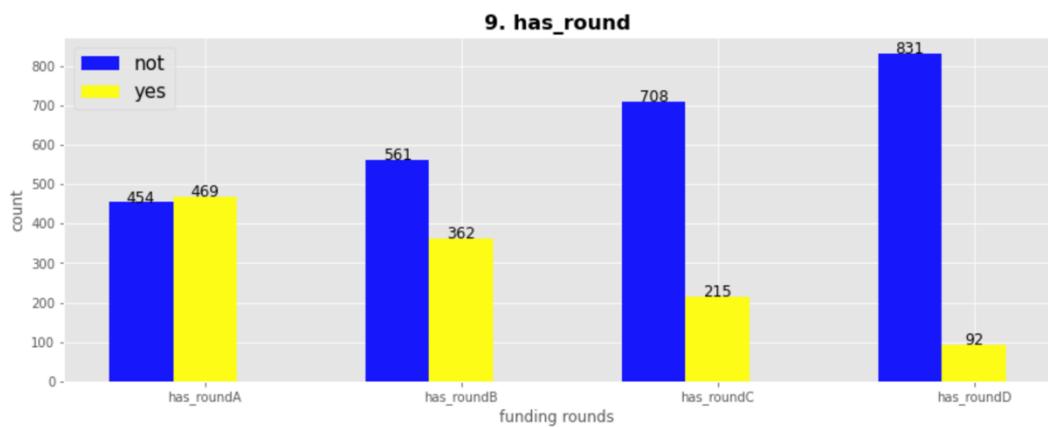


图 30 企业每轮集资是否进行分布图

从图 29、图 30 可以看出，融资轮数越多的企业数目越少，且越往后的轮次，越少企业进行。

3.2.6. 企业与外界的关系网

relationships 代表企业与外界建立的联系的数量。比如，企业与外界投资者、咨询师等有过联系，便算作是企业与外界建立的联系。

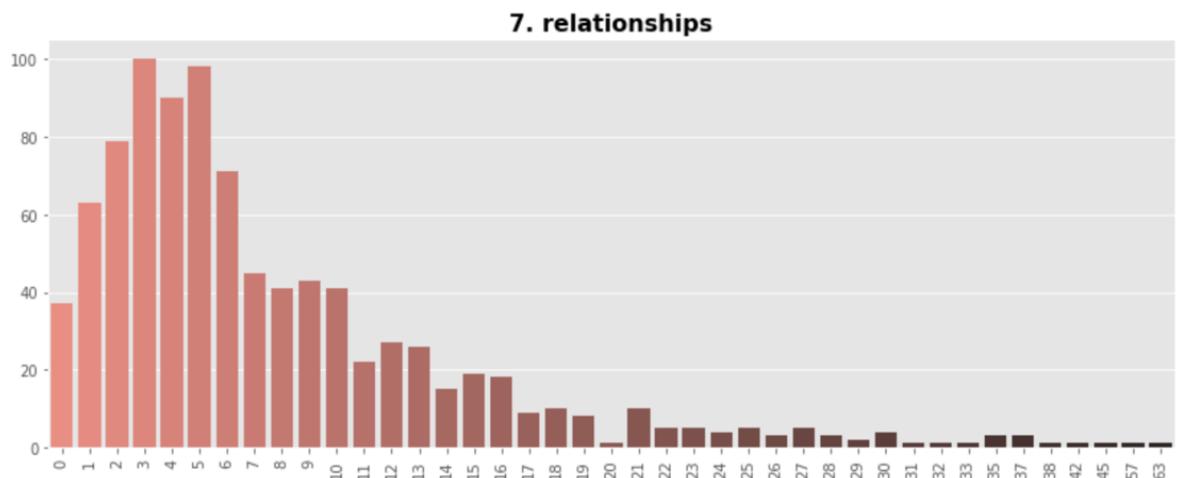


图 31 企业与外界关系网数量分布图

可以看出，大多数企业与外界的联系数目为 3-5 个。也有少量企业与外界联系网很庞大，这类企业可能是独角兽企业，也可能是行业本身的需求。

3.3. 数据预处理

在构建机器学习模型之前，首先对数据集进行预处理，以提高模型效果。

3.3.1. 删除无意义字段

根据字段的具体含义，我们删除了无法表征初创企业成功率特质的字段，包括无明确含义字段、重复含义字段、表征建立时间的字段、表示企业 id 的字段

| | | | | |
|------------|------------------|-----------------|-----------|---------------|
| Unnamed: 6 | Unnamed: 0 | closed_at | latitude | longitude |
| founded_at | first_funding_at | last_funding_at | status | founded_at_dt |
| zip_code | name | id | object_id | |

3.3.2. 删除质量较差的字段

删除缺失值较多和部分非 binary 且有重复含义的类别字段

| | | | |
|--------------------------|-------------------------|------------|--------------|
| age_first_milestone_year | age_last_milestone_year | state_code | state_code.1 |
|--------------------------|-------------------------|------------|--------------|

删除完字段后，对数据集做缺失值检视如下：

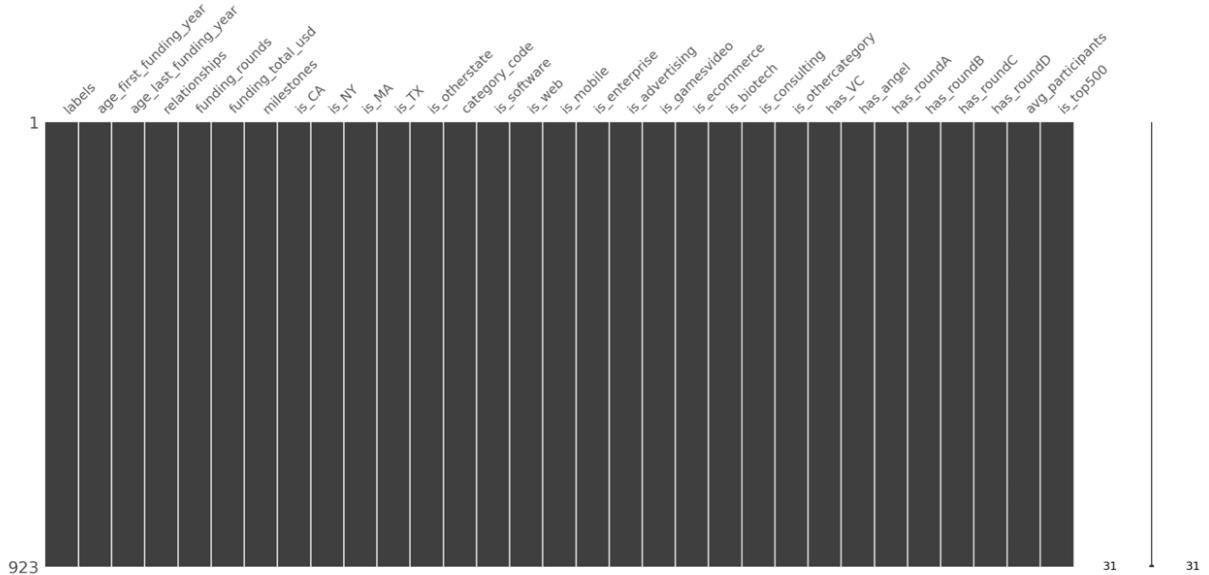


图 32 预处理后数据集缺失值分布

可以看出，已剔除所有缺失值。

3.3.3. 处理 category_code 字段

由于该字段为 multiple，且变量表示产业类别的英文名称，所以采取 get_dummies 的方法，将其转换为多个 0,1 取值的列，然后删除该字段。

转换后增加的字段：

| | | | | |
|----------------|-------------|----------------|---------------|------------------|
| advertising | analytics | automotive | biotech | cleantech |
| consulting | ecommerce | education | enterprise | fashion |
| finance | games_video | hardware | health | hospitality |
| manufacturing | medical | messaging | mobile | music |
| networkhosting | news | other | photo_video | public_relations |
| real_estate | search | security | semiconductor | social |
| software | sports | transportation | travel | web |

3. 4. 机器学习模型构建

为了预测初创企业的成功率，我们构建机器学习分类模型，根据企业的特征来评估企业的成功与否，从而为投资者提供投资方向。投资者可以根据本机器学习所采取的企业特征，将其调查的企业作为数据集，使用该机器学习模型对初创企业进行成功率的预测，以此为参考做出自己的投资决策。

3. 4. 1. 选取模型

先选取五个不同的模型进行建模。模型有如下 5 种：Random Forest Classifier、Light GBM Classifier、XGBoost Classifier、SVC、CatBoost Classifier。在此阶段，我们使用模型的默认参数（除了把分类对象设置为 binary 外）。

首先用 `train_test_split` 方法对数据集进行随机抽样，按 8:2 的比例将数据集划分为训练集和测试集。在训练集上分别对这五个未进行调参的模型进行训练，在测试集进行测试。各个模型的准确率如下：

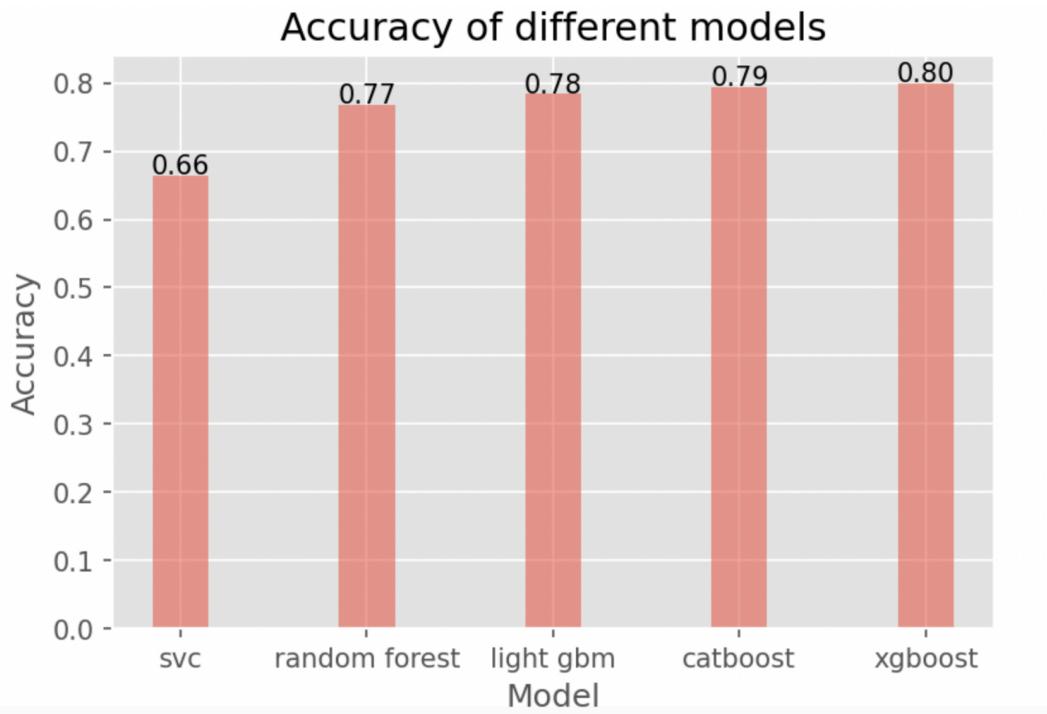


图 33 机器学习各模型准确率比较

之后，我们又使用 StratifiedKFold 方法进行分层取样、交叉验证，对每个模型求 accuracy 的均值。

在划分数据集的时候，为了尽可能保持数据分布的一致性，保持样本的类别比例相似，避免因数据划分过程引入额外的偏差而对最终结果产生影响，我们采用分层采样（stratified sampling）的手段对数据集进行划分。

为了避免过拟合现象，我们还采取交叉验证的方式对数据集进行测试。交叉验证是指，将数据集 D 划分为 k 个大小相似的互斥子集，然后，每次用 k-1 个子集的并集作为训练集，余下的那个子集作为测试集，这样就可获得 k 组训练/测试集，从而可进行 k 次训练和测试，最终返回这 k 个测试结果的均值^[7]。

StratifiedKFold 方法可以一步囊括分层取样和交叉验证的过程。我们取 5 折交叉验证，以 accuracy 作为评估目标，对五个机器学习模型做出评估如下：

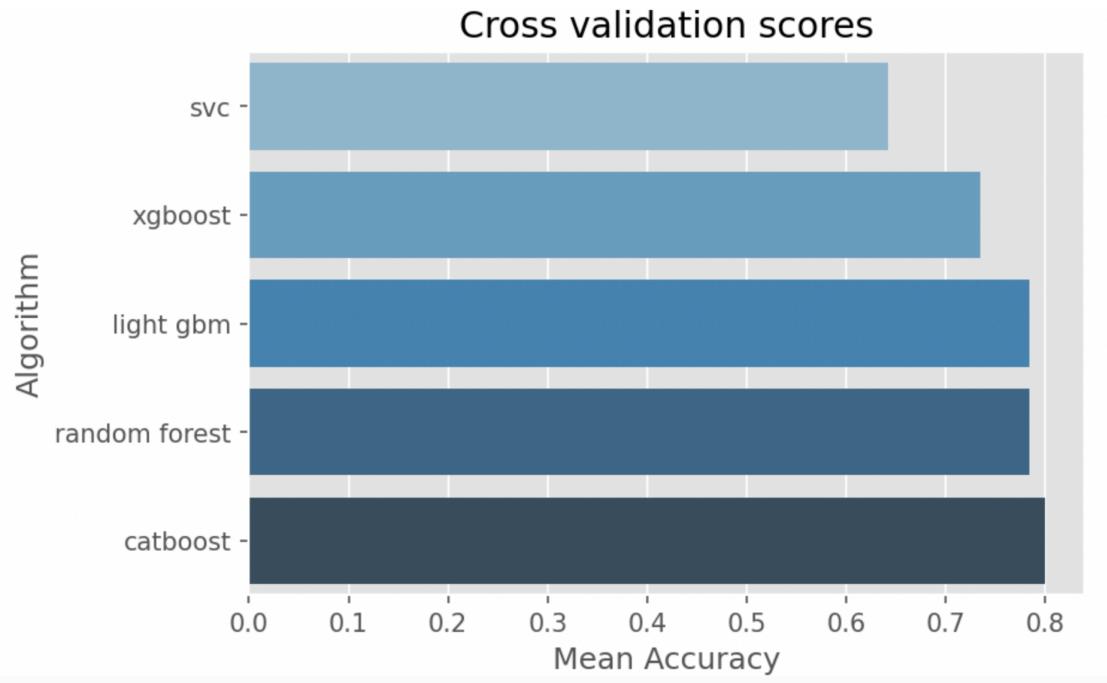


图 34 机器学习各模型交叉验证平均准确率比较

从上面图中可以看到，CatBoost 和 Random Forest 这两个模型，使用交叉验证评估的平均准确率最高。后续我们将选取这两个模型进行调参。

[7] 周志华，机器学习，清华大学出版社，2016

3.4.2. 调节超参数

我们使用 Bayes 调参对 Random Forest 模型调节超参数，使用 Optuna 对 Catboost 模型调节超参数，调参目标为优化 accuracy。

1. Bayes 调参中，我们选取 Random Forest 的三个重要超参数进行调参：

- n_estimators: number of trees in the forest
- max_feature: max number of features considered for splitting a node
- max_depth: max number of levels in each decision tree

调参过程如下：

| iter | target | max_depth | max_fe... | n_esti... |
|------|--------|-----------|-----------|-----------|
| 1 | 0.8303 | 5.919 | 0.7476 | 5.011 |
| 2 | 0.8455 | 5.116 | 0.2319 | 13.77 |
| 3 | 0.8474 | 4.304 | 0.4107 | 42.69 |
| 4 | 0.8529 | 6.772 | 0.4769 | 70.1 |
| 5 | 0.8406 | 4.431 | 0.8894 | 7.602 |
| 6 | 0.8535 | 9.035 | 0.4391 | 71.08 |
| 7 | 0.8523 | 6.451 | 0.5105 | 75.27 |
| 8 | 0.8396 | 10.0 | 0.9631 | 64.26 |
| 9 | 0.8377 | 3.0 | 0.999 | 72.6 |
| 10 | 0.8505 | 9.278 | 0.1548 | 74.06 |
| 11 | 0.8493 | 7.587 | 0.4438 | 78.62 |
| 12 | 0.8461 | 3.197 | 0.3317 | 80.91 |
| 13 | 0.8454 | 10.0 | 0.999 | 84.06 |
| 14 | 0.8444 | 8.989 | 0.1242 | 68.64 |
| 15 | 0.8448 | 7.341 | 0.999 | 72.44 |
| 16 | 0.8476 | 7.99 | 0.9318 | 76.22 |
| 17 | 0.8496 | 5.48 | 0.1737 | 76.78 |
| 18 | 0.846 | 9.985 | 0.7031 | 72.57 |
| 19 | 0.8519 | 7.946 | 0.1 | 70.58 |
| 20 | 0.8456 | 5.765 | 0.9688 | 68.49 |
| 21 | 0.854 | 6.928 | 0.695 | 81.03 |
| 22 | 0.8439 | 8.598 | 0.5586 | 80.92 |
| 23 | 0.8493 | 5.924 | 0.999 | 80.14 |
| 24 | 0.8551 | 6.185 | 0.4612 | 82.14 |
| 25 | 0.8433 | 6.845 | 0.999 | 83.31 |
| 26 | 0.8523 | 5.448 | 0.2266 | 81.58 |
| 27 | 0.8467 | 6.283 | 0.9839 | 81.34 |
| 28 | 0.843 | 6.604 | 0.2408 | 81.59 |
| 29 | 0.8487 | 7.189 | 0.2035 | 43.87 |
| 30 | 0.8409 | 8.874 | 0.4762 | 33.13 |

图 35 贝叶斯调参过程

调参后的最优参数为 {'max_depth': 6.927509259046254, 'max_features': 0.6950341039840612, 'n_estimators': 81.03428759971908}

2. Optuna 调参中，我们选取 Catboost 的三个重要超参数进行调参：

- depth: tree depth
- learning_rate: used for reducing the gradient step
- iterations: number of trees

调参过程如下：

```
[I 2022-06-17 13:15:33,908] A new study created in memory with name: no-name-50b50950-a5d3-43ed-894a-7b0d8049fa00
[I 2022-06-17 13:15:34,333] Trial 0 finished with value: 0.7783783783783784 and parameters: {'depth': 5, 'learning_rate': 0.18996237024652754, 'n_estimators': 5216}. Best is trial 0 with value: 0.7783783783783784.
[I 2022-06-17 13:15:35,048] Trial 1 finished with value: 0.772972972972973 and parameters: {'depth': 7, 'learning_rate': 0.2601528301721054, 'n_estimators': 5558}. Best is trial 0 with value: 0.7783783783783784.
[I 2022-06-17 13:15:45,137] Trial 2 finished with value: 0.7945945945945946 and parameters: {'depth': 10, 'learning_rate': 0.010142957084632018, 'n_estimators': 3224}. Best is trial 2 with value: 0.7945945945945946.
[I 2022-06-17 13:15:45,800] Trial 3 finished with value: 0.8054054054054054 and parameters: {'depth': 3, 'learning_rate': 0.03568355739936095, 'n_estimators': 5865}. Best is trial 3 with value: 0.8054054054054054.
[I 2022-06-17 13:15:47,477] Trial 4 finished with value: 0.8054054054054054 and parameters: {'depth': 6, 'learning_rate': 0.024245292221245697, 'n_estimators': 7597}. Best is trial 3 with value: 0.8054054054054054.
[I 2022-06-17 13:15:49,385] Trial 5 finished with value: 0.7783783783783784 and parameters: {'depth': 8, 'learning_rate': 0.056247309425964204, 'n_estimators': 6949}. Best is trial 3 with value: 0.8054054054054054.
[I 2022-06-17 13:15:51,091] Trial 6 finished with value: 0.7837837837837838 and parameters: {'depth': 9, 'learning_rate': 0.14834205369624792, 'n_estimators': 6475}. Best is trial 3 with value: 0.8054054054054054.
[I 2022-06-17 13:15:54,893] Trial 7 finished with value: 0.7945945945945946 and parameters: {'depth': 10, 'learning_rate': 0.0761448985563966, 'n_estimators': 4435}. Best is trial 3 with value: 0.8054054054054054.
[I 2022-06-17 13:15:55,734] Trial 8 finished with value: 0.8 and parameters: {'depth': 5, 'learning_rate': 0.17769768921778648, 'n_estimators': 3106}. Best is trial 3 with value: 0.8054054054054054.
[I 2022-06-17 13:15:58,325] Trial 9 finished with value: 0.8216216216216217 and parameters: {'depth': 10, 'learning_rate': 0.1215796508959345, 'n_estimators': 1943}. Best is trial 9 with value: 0.8216216216216217.
[I 2022-06-17 13:15:58,816] Trial 10 finished with value: 0.8108108108108109 and parameters: {'depth': 4, 'learning_rate': 0.09492388672560177, 'n_estimators': 9981}. Best is trial 9 with value: 0.8216216216216217.
[I 2022-06-17 13:15:59,195] Trial 11 finished with value: 0.8108108108108109 and parameters: {'depth': 3, 'learning_rate': 0.10754602958132012, 'n_estimators': 9962}. Best is trial 9 with value: 0.8216216216216217.
[I 2022-06-17 13:15:59,868] Trial 12 finished with value: 0.8162162162162162 and parameters: {'depth': 5, 'learning_rate': 0.09075251275218794, 'n_estimators': 2032}. Best is trial 9 with value: 0.8216216216216217.
[I 2022-06-17 13:16:01,051] Trial 13 finished with value: 0.8 and parameters: {'depth': 5, 'learning_rate': 0.0400434393213594, 'n_estimators': 1278}. Best is trial 9 with value: 0.8216216216216217.
[I 2022-06-17 13:16:02,483] Trial 14 finished with value: 0.7783783783783784 and parameters: {'depth': 8, 'learning_rate': 0.12058297458529209, 'n_estimators': 1197}. Best is trial 9 with value: 0.8216216216216217.
[I 2022-06-17 13:16:02,876] Trial 15 finished with value: 0.8054054054054054 and parameters: {'depth': 5, 'learning_rate': 0.2895359413309652, 'n_estimators': 2604}. Best is trial 9 with value: 0.8216216216216217.
[I 2022-06-17 13:16:03,631] Trial 16 finished with value: 0.7837837837837838 and parameters: {'depth': 6, 'learning_rate': 0.06943478853360383, 'n_estimators': 2171}. Best is trial 9 with value: 0.8216216216216217.
[I 2022-06-17 13:16:07,624] Trial 17 finished with value: 0.7891891891891892 and parameters: {'depth': 9, 'learning_rate': 0.02216044332403205, 'n_estimators': 4068}. Best is trial 9 with value: 0.8216216216216217.
[I 2022-06-17 13:16:08,551] Trial 18 finished with value: 0.8108108108108109 and parameters: {'depth': 4, 'learning_rate': 0.042724782529933, 'n_estimators': 2016}. Best is trial 9 with value: 0.8216216216216217.
[I 2022-06-17 13:16:10,481] Trial 19 finished with value: 0.8054054054054054 and parameters: {'depth': 8, 'learning_rate': 0.07990292527327578, 'n_estimators': 3842}. Best is trial 9 with value: 0.8216216216216217.
```

图 36 Optuna 调参过程

调参后的最优参数为 {'depth': 5, 'learning_rate': 0.026763419461744455, 'n_estimators': 6958}

我们将调参后的两个模型与之前未调参的五个模型再次进行交叉验证模型评估，结果如下：

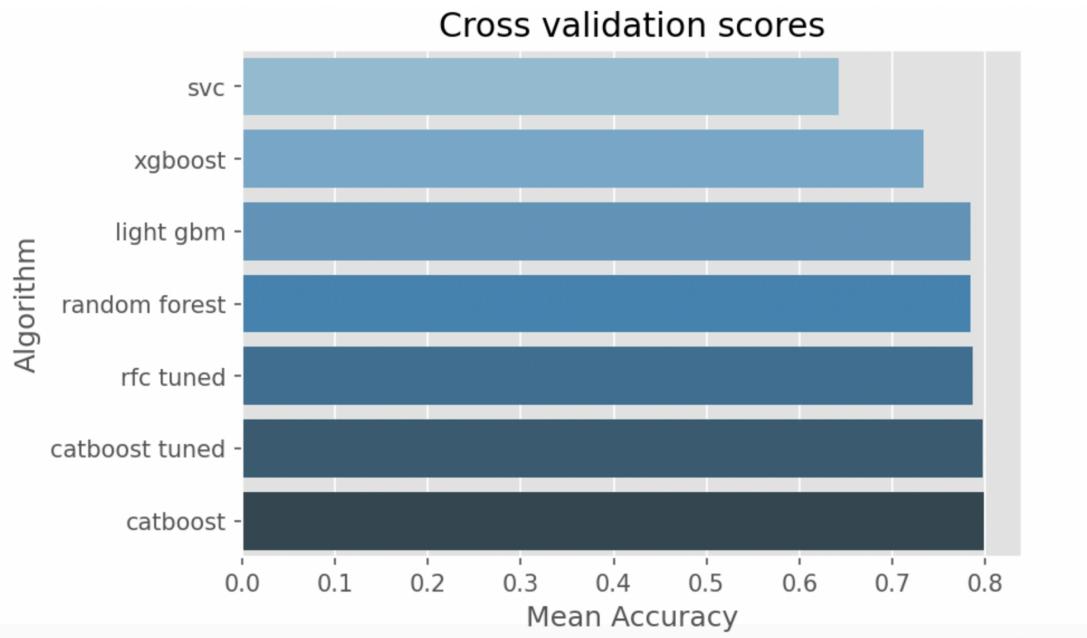


图 37 调参后模型与原模型做交叉验证平均准确率比较

可见，Random Forest 在进行调参之后的准确率有所提高。

未调参的 CatBoost 与调参后的相比，交叉验证准确率更高一些，这可能是因为我们将调参目标设置为 accuracy，未进行交叉验证，且优化目标单一，使得调参后模型存在过拟合现象。

由于 Catboost 和 Random Forest 采取的是不同的机器学习分类策略（前者为 boosted tree，后者为 random forest），所以在后续构建混合模型时，我们选择 boosted tree 中的最优模型，即未调参的 CatBoost，与 random forest 中的最优模型，即调参后的 Random Forest，这两个模型，以期得到最优效果的同时形成互补。

3.4.3. 建立集成模型

我们使用 sklearn 库中 Voting Classifier 的 soft voting 方法，对两个选取的模型进行集成，以期得到进一步的效果优化。

Voting Classifier 是一种 Ensemble Model，通过结合不同的机器学习分类模型，使用投票的方法来预测分类标签。这样的模型对于混合一组表现同样好但概念上不同的机器学习模型来说非常有用，可以平衡各自的弱点，使得模型效果达到更优。投票具体分为两种投票方式：一种是硬投票(Hard Voting)，即少数服从多数，输出被更多种分类模型预测得到的标签；一种是软投票(Soft Voting)，即计算平均概率，将每个分类模型预测标签的概率进行加权平均（权重自设），输出概率最大的标签。

由于本次只对两个分类模型进行混合，如果进行少数服从多数意味着二选一，信服力不够强，所以我们选择 Soft Voting 方法进行投票。

建立完集成模型后，对集成模型进行模型评估如下：

- 混淆矩阵

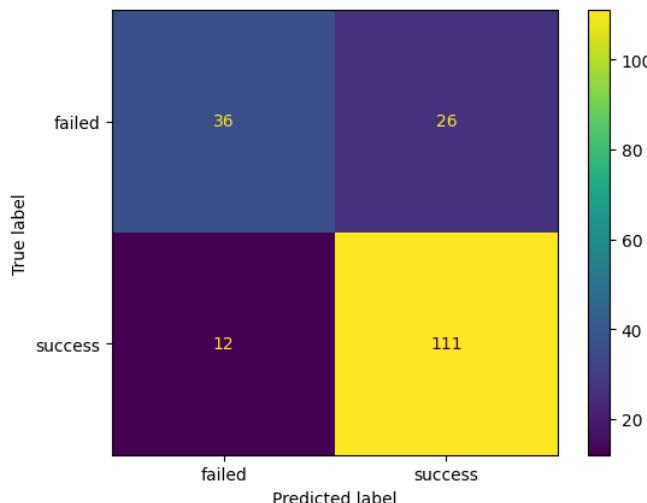


图 38 集成模型的混淆矩阵图

| 真实情况 | 预测情况 | |
|------|------|----|
| | 正例 | 反例 |
| 正例 | 111 | 12 |
| 反例 | 26 | 36 |

表 3 集成模型混淆矩阵表

- 模型效果评分

选取以下五个评分矩阵对模型的效果进行评估

$$(1) \text{accuracy} = \frac{TP+TN}{P+N}, \text{ 表示准确率}$$

$$(2) \text{precision} = \frac{TP}{TP+FP}, \text{ 表示精确率}$$

$$(3) \text{recall} = \frac{TP}{P}, \text{ 表示查准率}$$

$$(4) \text{F1 score} = \frac{2TP}{2TP+FP+FN}, \text{ 将精确率和查准率综合考虑, 范围[0,1]}$$

$$(5) \text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}, \text{ Matthews 相关系数, 范围[-1,1]}$$

MCC 用以测量二分类的分类性能的指标，该指标考虑了真阳性，真阴性，假阳性和假阴性，本质上是一个描述实际分类与预测分类之间的相关系数。它的取值范围为[-1,1]，取值为 1 时表示对受试对象的完美预测，取值为 0 时表示预测的结果还不如随机预测的结果，-1 是指预测分类和实际分类完全不一致。

| accuracy | precision | recall | f1 score | MCC score |
|----------|-----------|--------|----------|-----------|
| 0.795 | 0.810 | 0.902 | 0.854 | 0.520 |

表 4 集成模型效果评分表

| | accuracy |
|---------------------|----------|
| classifier | |
| SVC | 0.642 |
| XGBoost | 0.734 |
| lightgbm | 0.785 |
| random forest | 0.785 |
| Catboost | 0.799 |
| tuned random forest | 0.787 |
| tuned Catboost | 0.798 |
| voting classifier | 0.795 |

表 5 所有模型准确率比较

- ROC 曲线

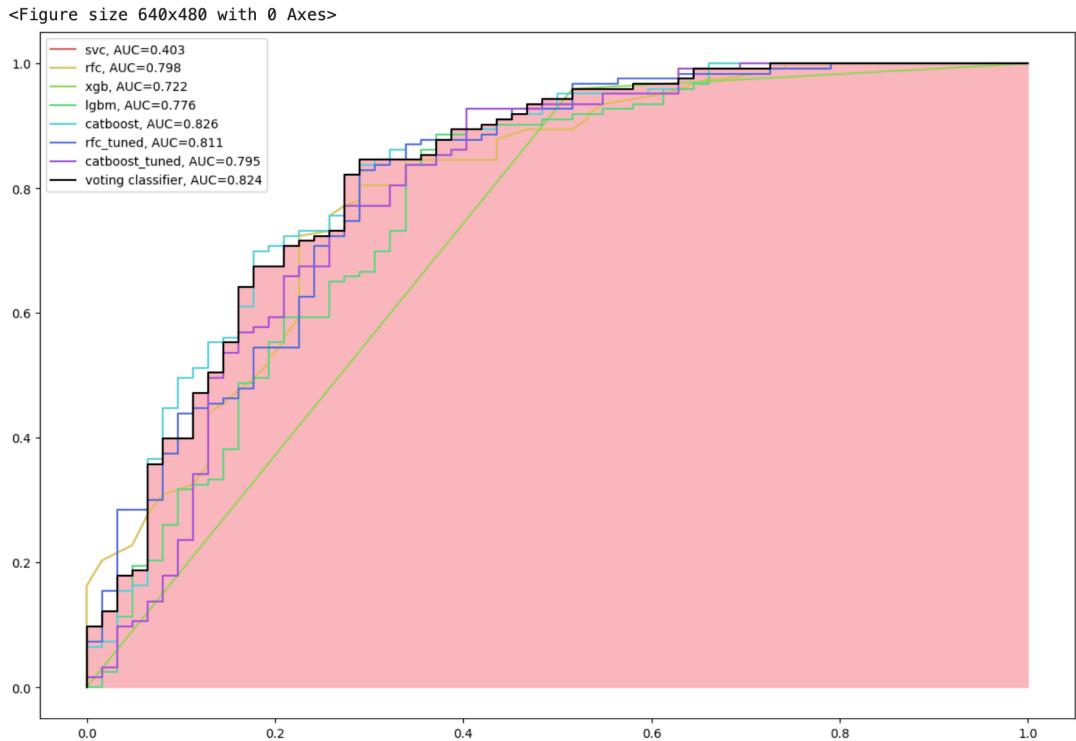


图 39 所有模型 ROC 曲线比较

比较 accuracy 和 ROC 曲线发现，集成学习模型的效果胜过其他模型，但是比未调参的 Catboost 略差。原因可能有以下几点：

- (1) Voting classifier 未进行调参，在求加权平均时 weights 默认为 None。比如，对于真实值为成功的数据，假如 Catboost 预测为成功的概率为 51%，tuned Random Forest 预测为成功的概率为 30%，则输出的结果为 $(51\% + 30\%) / 2 = 40.5\%$ ，即输出的标签为失败。在这种情况下，voting classifier 的准确率反而不如单模型的准确率高。
- (2) Voting classifier 中采用的模型之间的相关性会影响集成模型效果。我们之前对 random forest 进行的调参，是建立在单个模型的基础上进行的调参，而基于单个模型的调节得出的最优超参数，可能并不是使得集成模型最优的该模型的超参数。

然而，即使 voting classifier 的 accuracy 和 auc 比 Catboost 略低，我们还是决定采用 voting classifier 作为我们的最终模型。首先，集成学习的本质是通过平衡各模型的弱点从而达到更优的效果，目前效果欠佳可以通过调参进行进一步改善。其次，voting classifier 实际上增加了模型的超参数，使得模型优化提升的空间大大增加。所以，综合其基本理念和提升潜力，我们选择 voting classifier 作为我们本次作业的最终模型。

四、总结及反思

本次作业，我们使用 startup-analysis 与 startup-success-prediction 这两个数据集对初创企业的成功问题进行研究，两个数据集分别应用于分析企业成功因素与建立企业成功率预测模型，以期为创业者提供创业建议、为投资者提供投资方向。在分析企业成功因素的方向中，我们从创业者和创业公司两方面对成功企业的画像进行绘制，总结出成功企业的特征因素；在预测企业成功率的方向中，我们选取出较优模型进行超参数调节，并建立了集成模型以期达到更优的预测效果，最终模型准确率达到接近 80%，AUC 值也在 0.8 以上。

然而，我们目前的研究工作仍然有较大的改进空间：

- (1) 数据集方面，我们目前仅仅是在 Kaggle 网站进行了数据集检索。因此，我们的取样范围受到局限，地理位置主要集中于 USA；字段特征也并非全面，如关于 fundings 的特征字段较少，在第一个方向中未对 fundings 相关因素进行分析；同时，第一个数据集缺失值较多，在做因素分析时我们只是删除了对应字段的缺失值，未考虑其他字段，因此可能会引入样本选取的误差。后续改进方法有二：其一，我们可以从创业相关网站，如 crunchbase 等搜集更加全面的数据集进行研究；其二，我们可以引入额外具有相似特征、不同取样范围的数据集进行补充，以扩充样本数量。
- (2) 探索式数据分析方面，我们在处理特征字段时，未对时间序列相关特征进行分析，只是将整个时间跨度的样本综合到一起进行分析。后续我们可以按照不同的时间区间，对样本分别进行分析，纵向比较初创企业的发展问题，以期获得时间上变化发展的启发。
- (3) 机器学习预处理方面，对于质量不高的字段，我们只是删除了这些字段不予考虑，这样的操作可能会失去一些重要的特征信息。后续我们可以在预处理方面进行加强，选取合适的方法进行缺失值填充，以提高预测模型的普适性和有效性。
- (4) 机器学习模型构建方面，首先，我们在调参过程中存在调参后交叉验证准确率下降的问题，这可能是由于调参的优化目标仅仅为单折随机

划分的准确率，调参后反而出现过拟合现象。后续我们将通过多元化调优参数目标，以期得到更有效力的调参结果。其次，如 3.4.3 中分析，我们构建的集成模型的准确率和 AUC 反而略低于单个 Catboost 模型，这是因为我们未对集成模型设置集成权重，也可能是由于集成学习的两个模型相关性较强。后续我们将对集成模型的权重参数进行调优，如利用 GridSearchCV 等对模型进行调节参数，也会考虑选取相关性更弱的两个单模型进行集成学习。

参考文献：

- [1] Rebecca Baldridge, and Benjamin Curry. 2021. What is a Startup? *Forbes* (April).
<https://www.forbes.com/advisor/investing/what-is-a-startup/>
- [2] Ries, E. (2011). The Lean Startup. *Working Paper*, 1–28. <http://doi.org/23>
- [3] Hanley, J. A., & McNeil, B. J. (1982). The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. *Radiology*, (143). Retrieved from <http://pubs.rsna.org/doi/pdf/10.1148/radiology.143.1.7063747>
- [4] Kim, E. (2015). Fastest startups to \$1 billion valuation - Business Insider.
Retrieved August 21, 2017, from <http://www.businessinsider.com/fastest-startups-to-1-billion-valuation-2015-8/#1-slack-is-the-fastest-growing-enterprise-software-ever-11111114>
- [5] Francisco Ramadas. 2017. Predicting Start-up Success with Machine Learning.
University of Lisbon. Published in November
<https://run.unl.pt/bitstream/10362/33785/1/TGI0132.pdf>
- [6] Anna Melnychuk. 2021. STARTUP SUCCESS PREDICTION. EXAMPLE FROM THE US. <https://kse.ua/wp-content/uploads/2021/12/Anna-Melnychuk-1>
- [7] 周志华, 机器学习, 清华大学出版社, 2016