企业获投潜力评估

# 背景

判断企业的发展前景会有各种各样的要素禀赋，诸如资本实力、知识产权、人员规模结构、所在行业的发展趋势等等，而投资机构灵敏的嗅觉亦是重要的先行指标，无论是资本市场广受追捧的佼佼者还是独角兽企业，在每家企业的发展历程中，基本上都能找到投资机构的影子。因此，对于一家企业（尤其是创业企业）能否获得投资机构的青睐，是跻身“好企业”的重要风向。

杭州索骥数据科技有限公司旨在对标最优秀的中国企业，目前已开源众多优质企业准入门槛，共收录符合准入门槛的优质企业逾50万家。本课题通过企业几十个维度标签画像数据的提供，从不同维度对企业获得投资的潜力进行分析和预测，对于政府部门的招商、金融机构的项目发掘有着极其重要的意义。

# 任务描述

## 数据提供方

杭州索骥数据科技有限公司

## 数据说明

本赛题提供企业基本信息、股东信息、获投信息、专利信息、裁判文书信息、招聘信息等，包括如下字段：

|  |  |  |
| --- | --- | --- |
| enterprise\_id | 企业ID | 抽样&脱敏 |
| invest\_status | 2016.6.30～2017.6.30是否获投 | 1、是；0、否（只有样本数据中有此字段） |
| registered\_time | 企业注册时间 | yyyy/mm/dd |
| registered\_capital | 企业注册资本 | 整型值（单位万元） |
| industry\_code | 企业所属行业 | 国标一级代码 |
| Address | 企业地址 | 文字描述 |
| Tag | 企业科技标签 | 1、国家千人计划，2、国家高新技术企业，3、省科技型企业，4、市高新企业 |
| Product | 企业产品简介 | 文字描述 |
| employees\_num | 企业参保人数 | 整型值 |

|  |  |  |
| --- | --- | --- |
| enterprise\_id | 企业ID | 抽样&脱敏 |
| shareholder\_name | 股东名 | 文字描述 |
| shareholder\_type | 股东类型 | 1、个人，2、实体企业，3、基金 |
| shareholder\_percent | 持股比例 | 0～100 |

|  |  |  |
| --- | --- | --- |
| enterprise\_id | 企业ID | 抽样&脱敏 |
| invest\_name | 投资机构名称 | 脱敏 |
| invest\_id | 投资机构id | 抽样&脱敏 |
| invest\_time | 投资时间 | yyyy/mm/dd |
| invest\_level | 投资机构实力 | 1、大型，2、中型，3、小型，4、微型 |

|  |  |  |
| --- | --- | --- |
| enterprise\_id | 企业ID | 抽样&脱敏 |
| patent\_name | 知识产权名称 | 抽样&脱敏 |
| patent\_desc | 知识产权摘要 | 抽样&脱敏 |
| patent\_time | 知识产权时间 | yyyy/mm/dd |
| patent\_type | 知识产权类型 | 1、发明专利，2、实用新型，3、外观专利 |

|  |  |  |
| --- | --- | --- |
| enterprise\_id | 企业ID | 抽样&脱敏 |
| case\_name | 案件名称 | 抽样&脱敏 |
| case\_time | 裁判日期 | yyyy/mm/dd |
| case\_type | 案件类型 | 1、刑事案件，2、民事案件，3、行政案件，4、赔偿案件，5、执行案件，6、知识产权，7、其他类型 |

## 任务描述

### 训练数据部分

本赛题提供2016年6月30前的相关数据，选手从数据中自行组成特征和数据格式，自由组合训练测试数据比例。

### 预测数据部分

### 预测这些企业在2016年6月30日至2017年6月30日之间获得新一轮投资的概率。

数据中部分列存在空值或NULL，请参赛者自行处理。

为保护企业的隐私和数据安全，所有数据均已进行了采样和脱敏。

### 评分数据格式

参赛者完成预测后，需要将结果放入指定格式的数据表中，并去重，表字段要求如下：

|  |  |  |
| --- | --- | --- |
| enterprise\_id | 企业ID |  |
| invest\_probability | 获投概率 | 0～1 |

### 初赛数据

初赛阶段提供8000家企业数据。

### 决赛数据

决赛阶段提供5万家企业数据。

## 评估指标

本赛题采用以下维度进行评估。具体如下：

ROC/AUC

AUC =

其中，为第i条样本的序号（概率分从大到小排，排在第rank位置），M，N为获投企业个数和未获投企业个数。我们以AUC值作为最终的唯一评判标准。