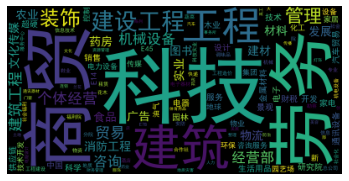
# 一、模型建立

## 1、数据分析

### 行业分类及行业规模

根据分词得到各个公司的行业关键词，然后根据行业关键词来判断其属于哪些行业。



* 1. 图1. 分词结果

得到行业之后，根据可能出现的词语分类，以下是分类标准：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 标签 | 类别 | 关键词 | 中型企业 | 小型企业 | 微型企业 |
| 0 | 租赁和商务服务业 | 商贸、劳务、机械设备、贸易、管理、咨询 | ＞8000万 | >100万 | <100万 |
| 1 | 工业 | 装饰、工程、建筑、建设工程、建筑工程 | ＞2000万 | >300万 | <300万 |
| 2 | 软件和信息技术服务业 | 软件和信息技术服务业 | >1000万 | >50万 | <50万 |
| 3 | 其他 | —— | >1000万 | >50万 | <50万 |

根据上表来划分行业类别和行业规模，分别以3、2、1对应中、小、微。

### 企业进项/销项总金额

企业销项金额 = 有效进项/销项发票**金额**之和（这里用数学公式表示）

### 企业总税额

企业总税额 = 有效进项发票税额之和-有效销项发票税额之和（这里用数学公式表示）

### 企业进项/销项数

企业销项数 = 企业有效进项/销项发票数

### 作废发票数

作废发票数 = 企业作废发票数//企业进项/销项发票数

### 作废比例

作废比例 = 作废发票数// (企业进项数+企业销项数)

### 进项(销项)年(月)数

进项(销项)年(月)数 = 企业进项(销项)发票出现过的年(月)份总数

### 供求关系

供应公司比重 = 企业有效进项发票出现过的公司总数/企业有效进项发票数

销售公司比重 = 企业有效销项发票出现过的公司总数/企业有效销项发票数

## 2、极限随机树（Extra Trees）

**随机森林：**是一个包含多个决策树的分类器， 并且其输出的类别是由个别树输出的类别的众数而定。，随机森林对回归的结果在内部是取得平均但是并不是所有的回归都是取的平均，有些是取的和。

* 样本随机
* 特征随机
* 参数随机
* 模型随机（ID3 ,C4.5）

**极限树/极端随机树：** 与随机森林算法十分相似，都是由许多决策树构成。

* 特征随机
* 参数随机
* 模型随机（ID3 ,C4.5）
* 分裂随机

**主要区别：**

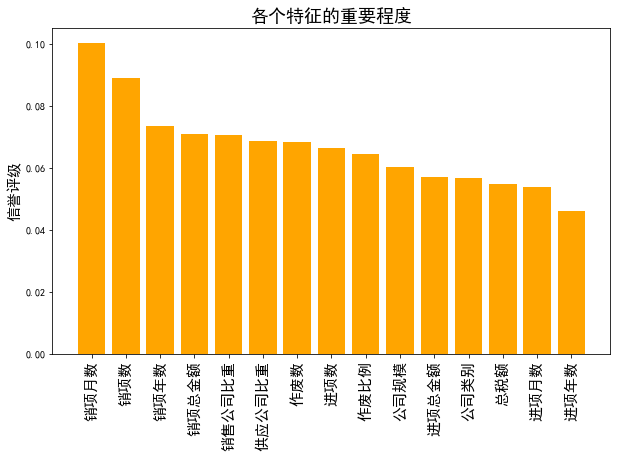
1. 极限树应用的是Bagging模型,extraTree使用的所有的样本，只是特征是随机选取的，因为分裂是随机的，所以在某种程度上比随机森林得到的结果更加好
2. 随机森林是在一个随机子集内得到最佳分叉属性，而ET是完全随机的得到分叉值，从而实现对决策树进行分叉的。

我们仅以二叉树为例，当特征属性是类别的形式时，随机选择具有某些类别的样本为左分支，而把具有其他类别的样本作为右分支；当特征属性是数值的形式时，随机选择一个处于该特征属性的最大值和最小值之间的任意数，当样本的该特征属性值大于该值时，作为左分支，当小于该值时，作为右分支。这样就实现了在该特征属性下把样本随机分配到两个分支上的目的。然后计算此时的分叉值（如果特征属性是类别的形式，可以应用基尼指数；如果特征属性是数值的形式，可以应用均方误差）。遍历节点内的所有特征属性，按上述方法得到所有特征属性的分叉值，我们选择分叉值最大的那种形式实现对该节点的分叉。从上面的介绍可以看出，这种方法比随机森林的随机性更强。

对于某棵决策树，由于它的最佳分叉属性是随机选择的，因此用它的预测结果往往是不准确的，但多棵决策树组合在一起，就可以达到很好的预测效果。

当ET构建好了以后，我们也可以应用全部的训练样本来得到该ET的预测误差。这是因为尽管构建决策树和预测应用的是同一个训练样本集，但由于最佳分叉属性是随机选择的，所以我们仍然会得到完全不同的预测结果，用该预测结果就可以与样本的真实响应值比较，从而得到预测误差。如果与随机森林相类比的话，在ET中，全部训练样本都是OOB样本，所以计算ET的预测误差，也就是计算这个OOB误差。

## 模型结果



3. 图1 信誉评级和各个特征重要程度排序

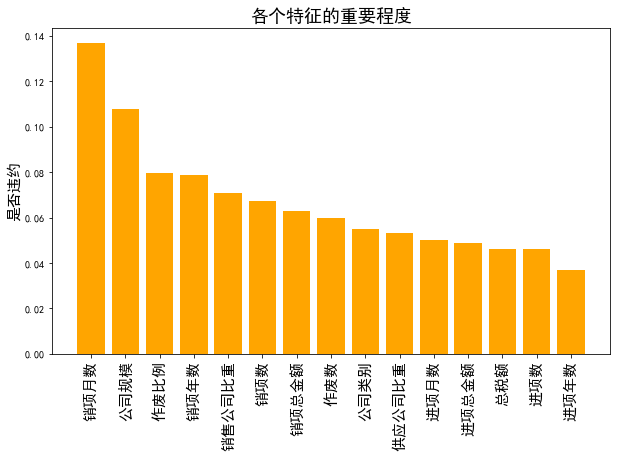
1. 销项月数 0.100146
2. 销项数 0.089040
3. 销项年数 0.073568
4. 销项总金额 0.070872
5. 销售公司比重 0.070529
6. 供应公司比重 0.068767
7. 作废数 0.068216
8. 进项数 0.066374
9. 作废比例 0.064396
10. 公司规模 0.060228
11. 进项总金额 0.056994
12. 公司类别 0.056592
13. 总税额 0.054738
14. 进项月数 0.053617
15. 进项年数 0.045924

模型准确率 0.8864864864864865

模型精确率 0.8956349206349206

模型召回率 0.9263070263070263

模型F1 0.88975322922691344



3. 图2 是否违约和各个特征重要程度排序

1. 销项月数 0.136774
2. 公司规模 0.107662
3. 作废比例 0.079521
4. 销项年数 0.078600
5. 销售公司比重 0.070696
6. 销项数 0.067422
7. 销项总金额 0.062892
8. 作废数 0.059975
9. 公司类别 0.054940
10. 供应公司比重 0.053284
11. 进项月数 0.050241
12. 进项总金额 0.048988
13. 总税额 0.046144
14. 进项数 0.046077
15. 进项年数 0.036785

模型准确率 0.9648648648648649

模型精确率 0.9375

模型召回率 0.8327586206896552

模型F1 0.8667087011349306

# 二、解题思路

## 确定信誉等级和是否违约

附件一中已有信誉等级和违约数据，但是我们可以上述模型获得可能违约的公司，故而应用上述的违约模型获取对应的**风险等级**（0-无风险，1-有风险），对于附件二，则同时应用信誉等级模型和违约模型获得信誉等级和是否违约。

## 数据筛选

根据上述的信息，我们将以下几类的公司排除在贷款考虑之外：

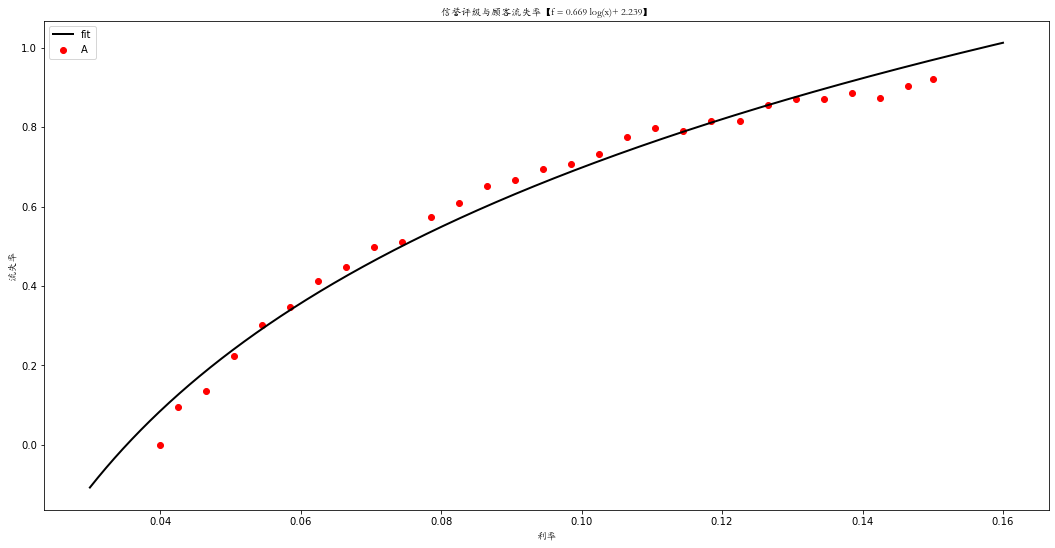
* 信用为D
* 已经违约
* 风险等级为1（可能违约）
* 企业总税额为负数

这样我们就得到了一份可以贷款的企业名单，按照

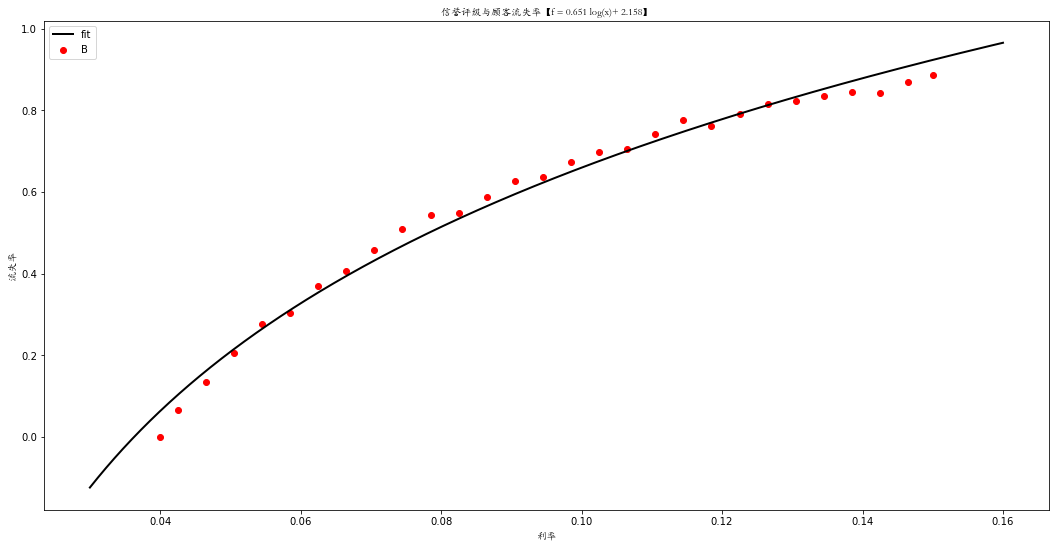
'信誉评级','总税额','公司规模','作废比例'

来进行排序信誉等级越高，总税额越多，公司规模越大，作废比例越小。

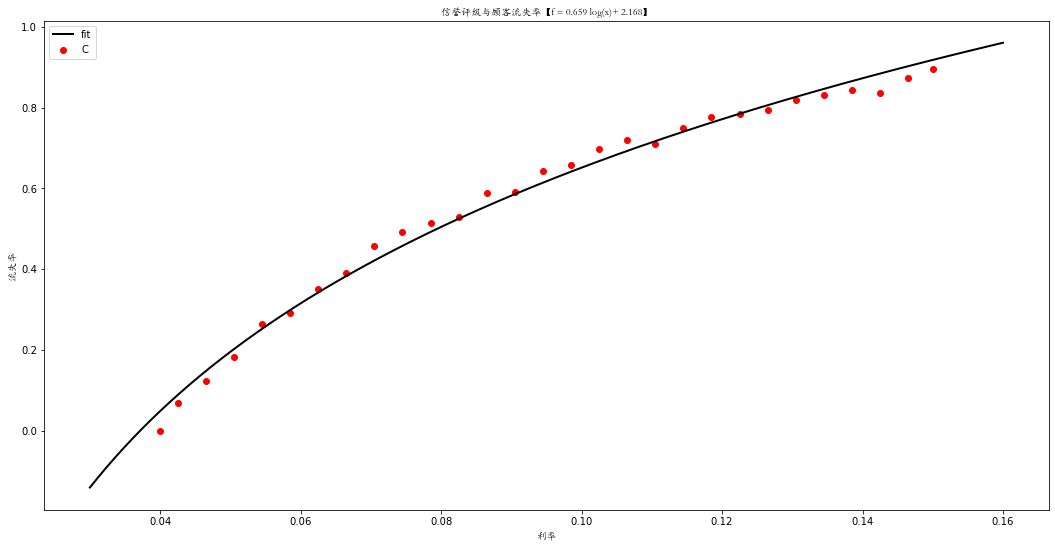
## 税率和客户流失率关系



二-3 图1 A信誉客户税率和流失率散点图及对数回归结果



二-3 图2 B信誉客户税率和流失率散点图及对数回归结果



二-3 图3 C信誉客户税率和流失率散点图及对数回归结果

A: 0.6690026022949184\*log(x) + 2.238599821505864

B: 0.65057086531514\*log(x) + 2.1576334615275665

C: 0.6593787589707425\*log(x) + 2.188076580002009

这个用数学公式表达

简介：<https://www.sohu.com/a/243223422_617676>

对数回归模型，建议详尽解释

## 供求稳定性与税率

供求越稳定，税率越低【最好找点啥来证明支持一下这个】

供求关系在上面表述过了，可以回去看一下。

将供求区间映射到税率区间内，映射方法最好用图表示一下。

公式和术语表示

## 税率期望

上一步骤获得了两个税率，依照两个税率结合步骤3中获得的函数，求得流失率。

期望=企业总税额\*税率\*(1-流失率)

获得两个期望，取最大的期望以及对应的税率。

## 贷款额度

贷款额度在总额不确定下，按照我们的排序，依照

额度=月均进项款\*（1-作废比例）\*1.1

以万为单位发放。

在总额确定情况下，根据上述各种贷款额度

额度=月均进项款\*（1-作废比例）

按照排序发放直到贷款数目没有了。