



# 互联网金融风控模型

刘时斌  
@数信互融

## 刘时斌

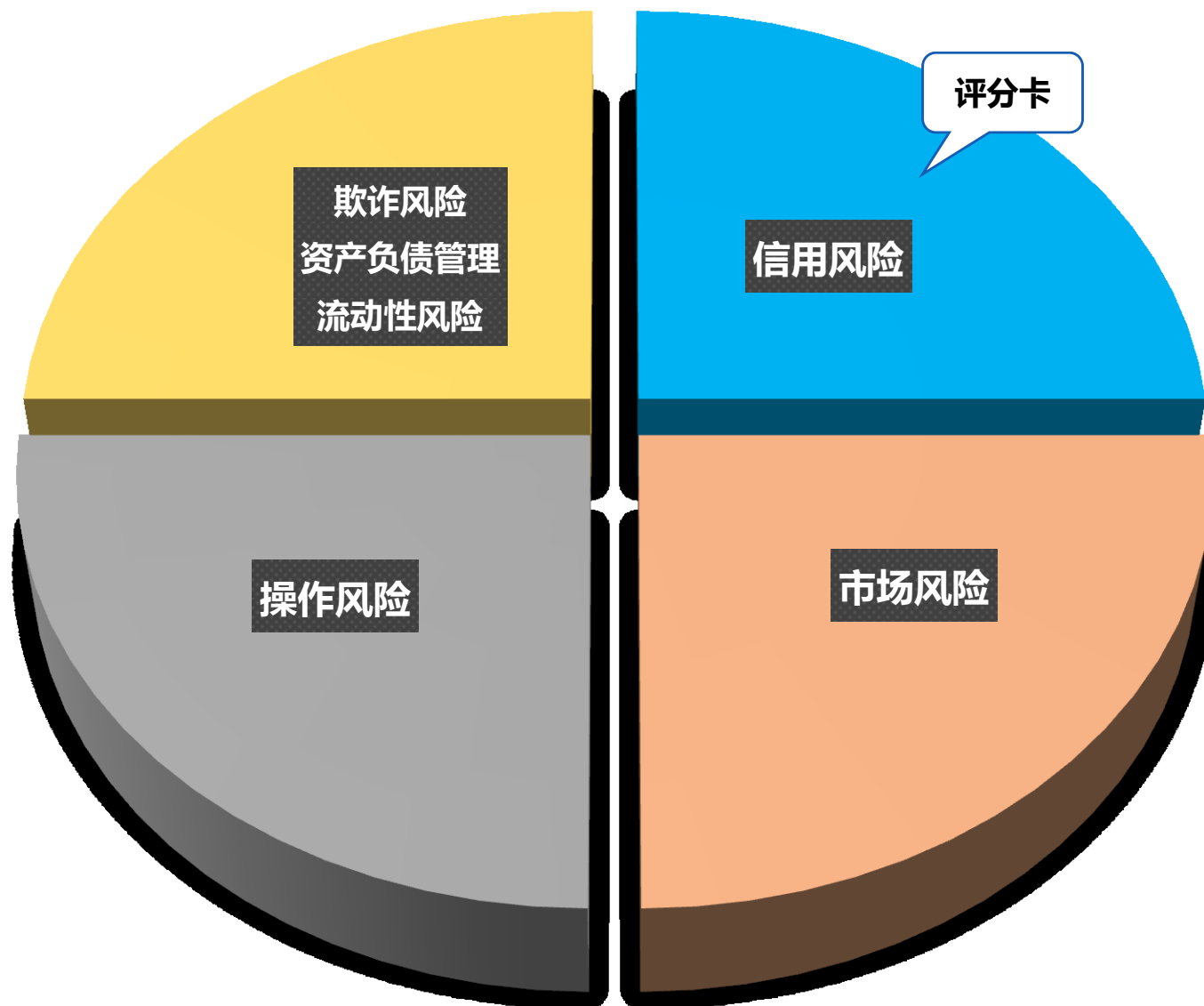


- 数信互融研发负责人，联合创始人
- 统计学硕士。法国INSA de Toulouse
- 曾在SAS任职负责风险产品开发超过十年、拥有丰富的风险产品开发经验
- 作为SAS中国和北京大学战略合作项目实施人、在北京大学连续三年主讲“统计分析和商务智能”课程（48学时、3学分），培养北大学生超过200人。

- 邮箱：[liushibin@ifre.com.cn](mailto:liushibin@ifre.com.cn)
- 微信：liushibin\_qd
- 手机：13910062734



# 风险类别-BASEL



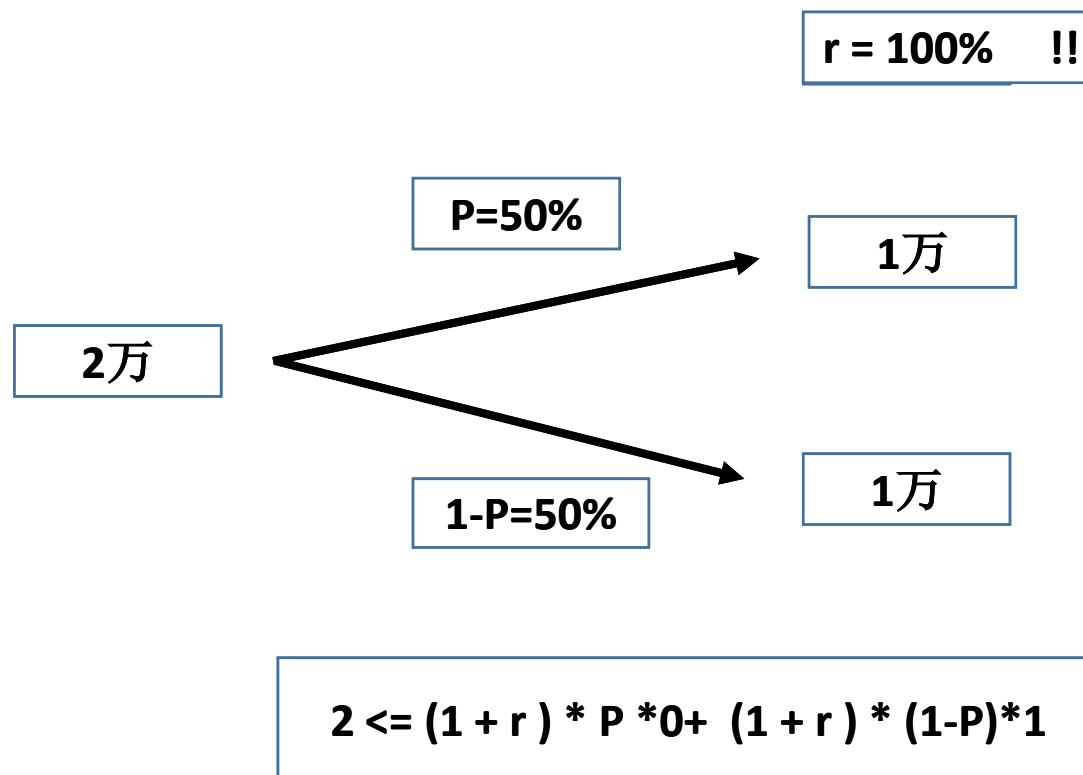
# 风险管理：零风险？

**RAROC ( Risk Adjusted Return on Capital )**  
即风险调整资本收益  
平衡收益和风险，即收益/风险,承担每单位风险的  
基础上收益的最大化.



**风险管理**

# 金融的本质：风险定价



# 金融的本质： 风险定价

**$r = 100\%$  !!**

监管规则

违约率评估

行业竞争水平

# 我们的模型服务对象

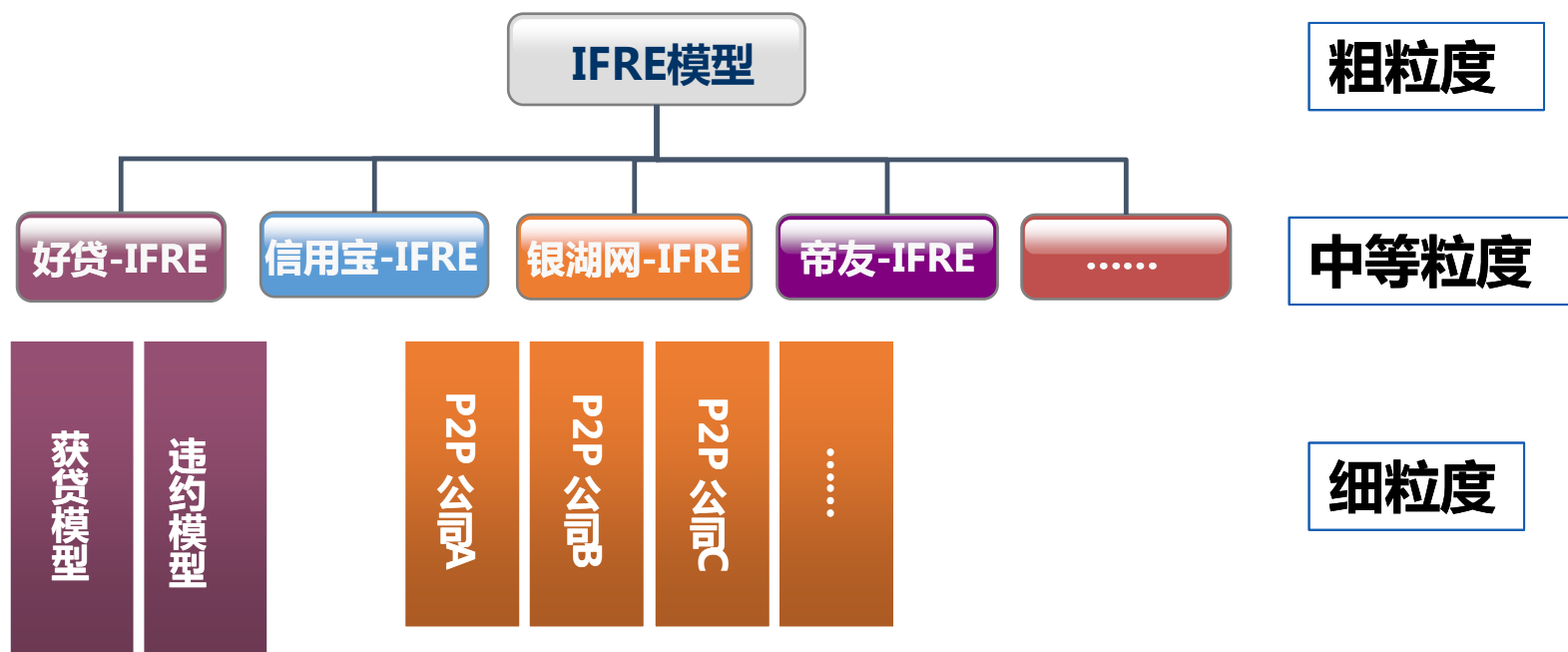
小微型企业信用评估

德国  
IPC

个人/消费者信用评  
估

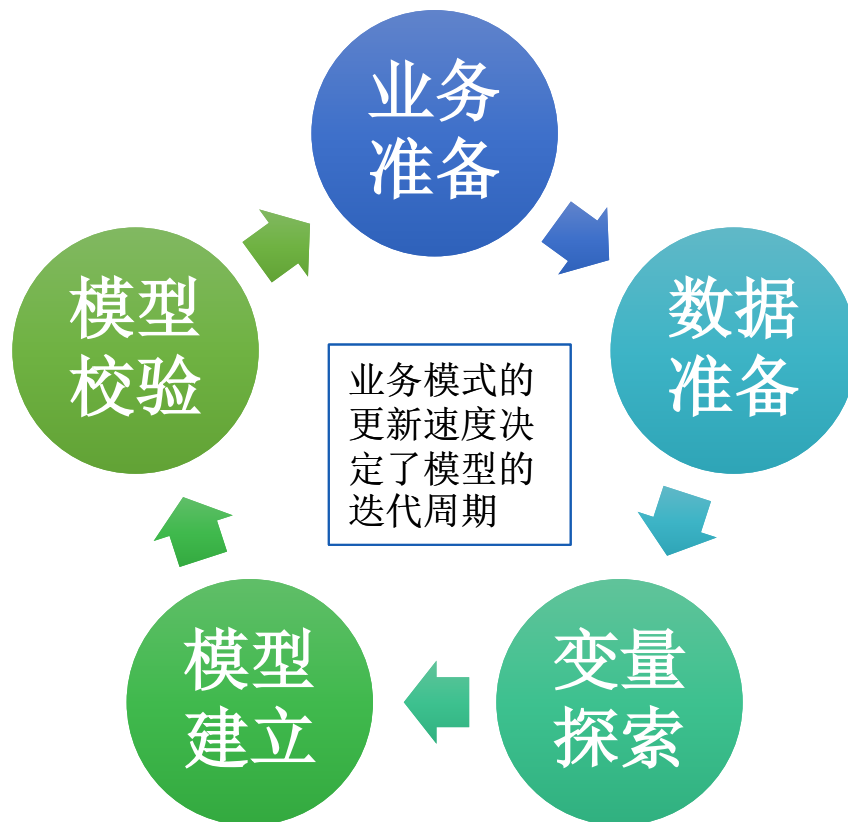
FICO

# 模型的适用性

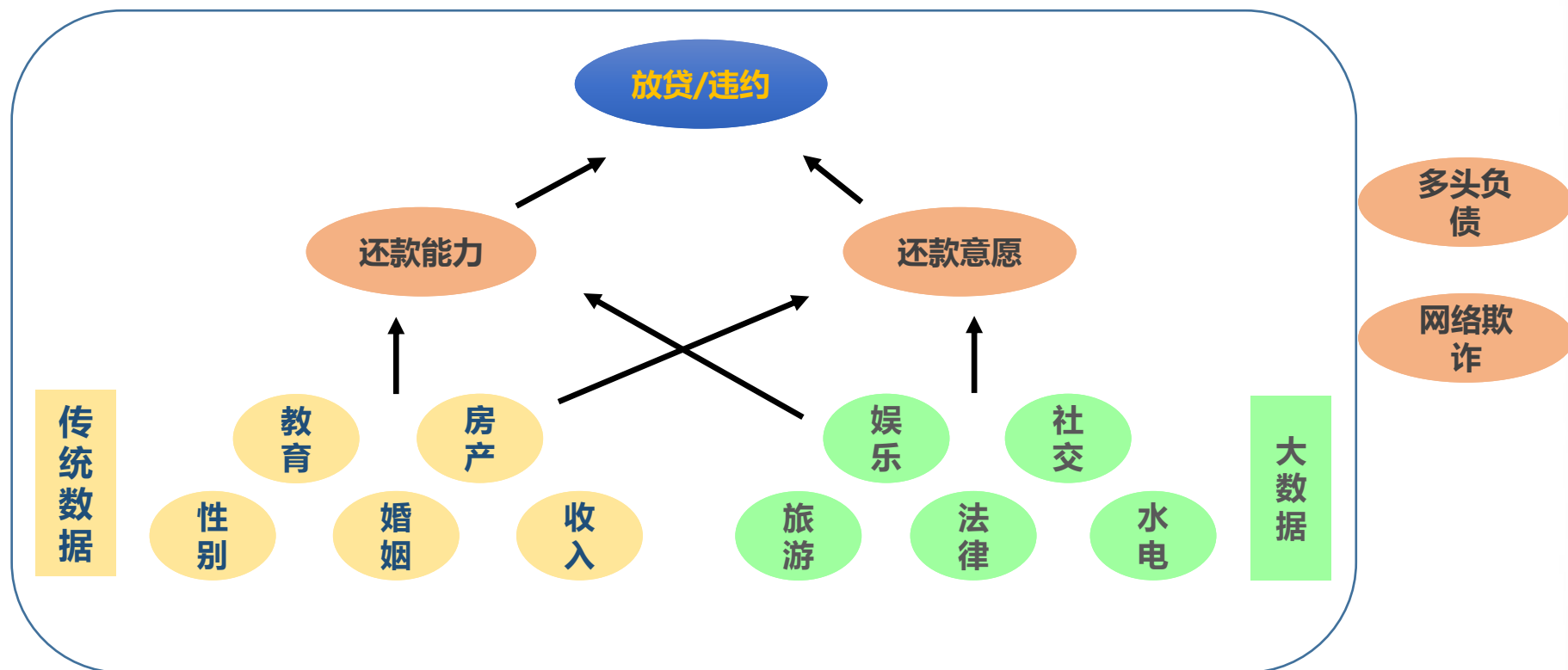


# 模型开发周期

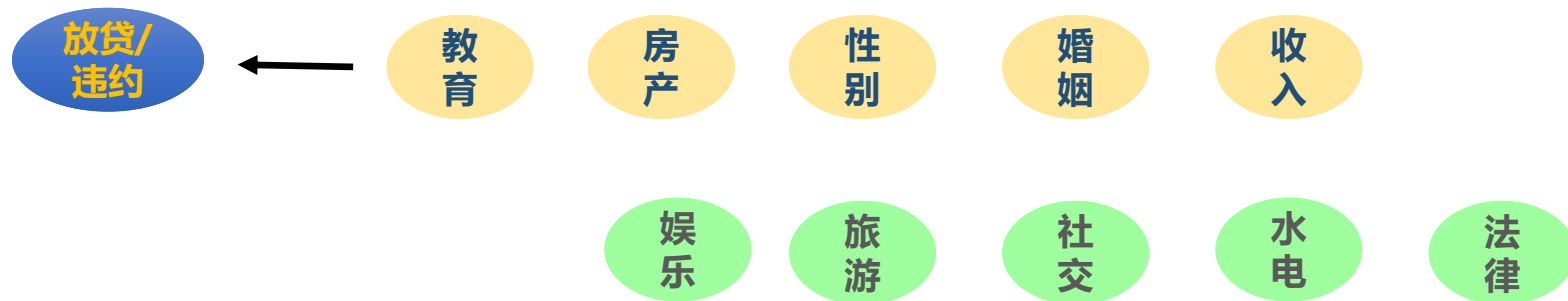
模型生命周期



# 算法：有监督的学习



# 传统数据和大数据的拼接



# 大数据对模型和风控的影响

- 非结构化数据：影音，图像，文本
  - 存储，传输，运算
  - 数据维度的增加，即第三方数据的接入
  - 行为数据
- 
- 传统结构化数据和非结构化数据的投入产出比：8/2 ？

# 大数据对模型和风控的影响

- 数据是不是越多越好？
  - 有代表性 ( Representative )
  - 数据中 所从呈现出的规律是否稳定 ( Stationary )

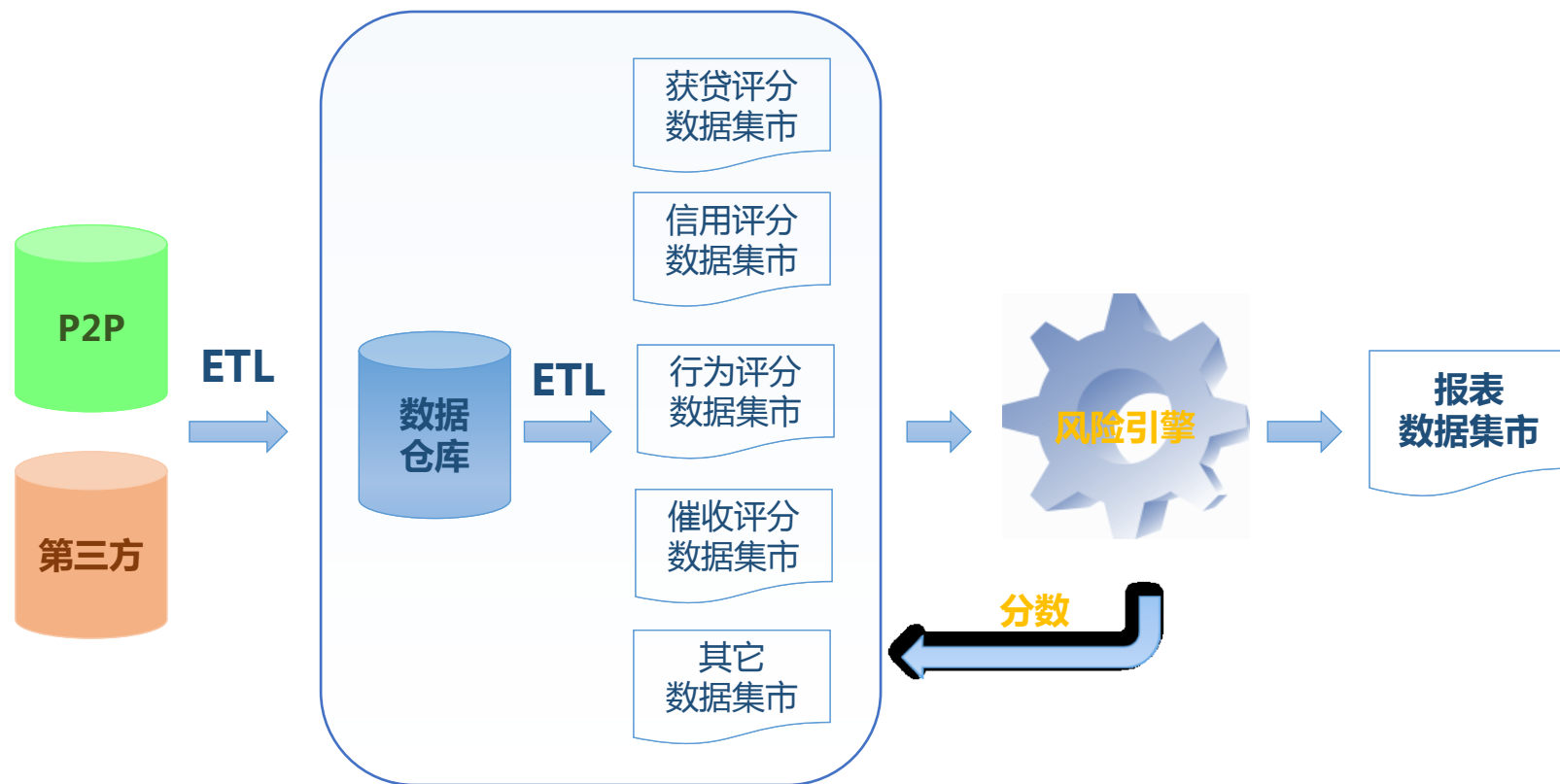


# 最少需要多少数据？

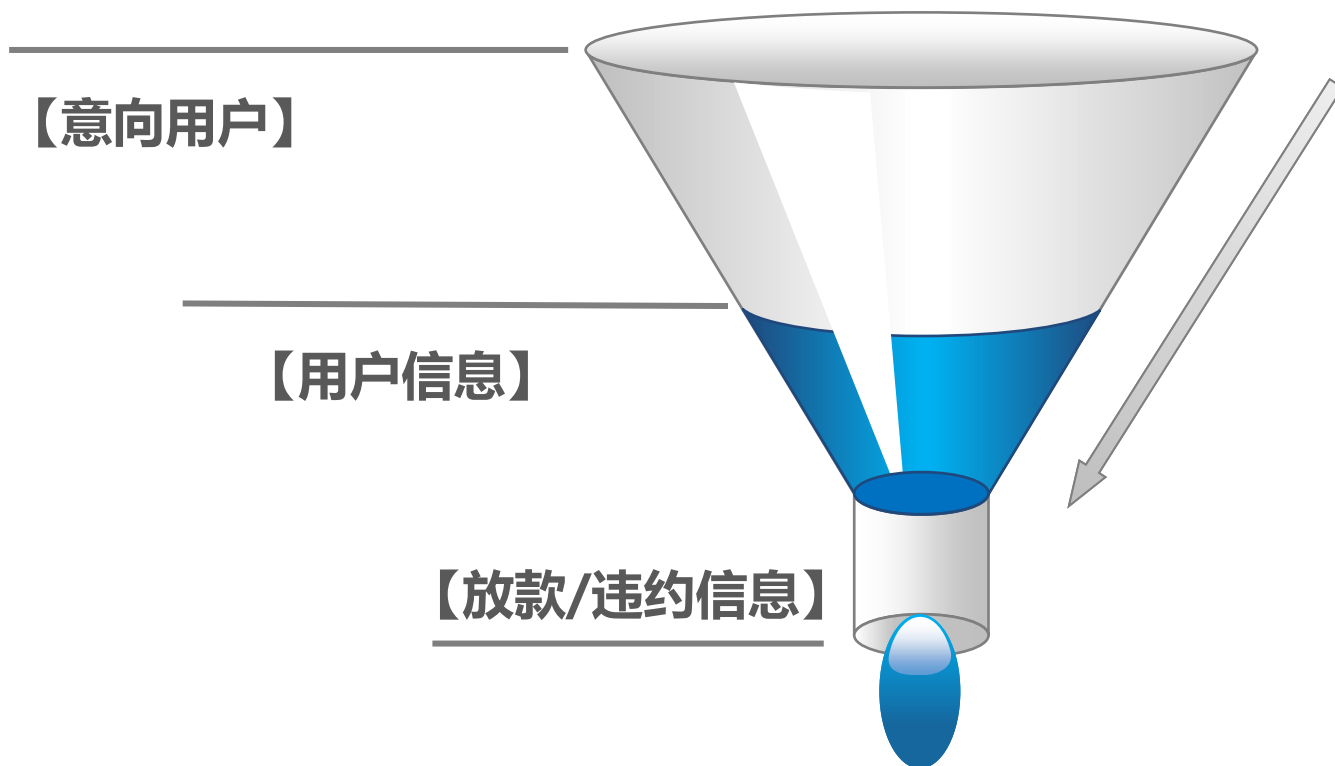


如果有10个变量，每个有2个属性值。  
则需要 $2^{10}=1024$  条数据才能覆盖  
所有属性的组合。

# 数据入库---数据仓库



# 数据清洗



## 数据准备：图像数据结构化

ID	注册号	名称	住所	法人	公司类型	经营范围	注册资本（万）	
1	111111	XXXXXX	XXXXXX	张XX	个体	服装	8700	
				实收资本（万）	成立日期	营业期限_起	营业期限_止	审核
				8700	2010.01.01	2010.01.01	2020.01.01	通过

企 业 法 人 营 业 执 照		注 册 号	11111111111111
名 称	XXXXXXXXXXXXXXXXXXXX	注 册 资 本	8700万元
住 所	XXXXXXXXXXXXXXXX	实 收 资 本	8700万元
法定代表人姓名	XXX		
公 司 类 型	其他有限责任公司		
经 营 范 围	许可经营项目：无 一般经营项目：XXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXX		
成 立 日 期	2001年02月22日	2010 年 1 月 20 日	
营 业 期 限	自 2001年02月22日 至 2051年02月21日		

请于每年3月1日至6月30日向登记机关申报年检

中华人民共和国国家工商行政管理总局制

## 数据准备：文本数据结构化

id	reviews
387	本地人，25周岁，客户在中山私营自己做生意，营业执照注册3年，是法人，每月流水20-30万，信用记录良好，之前在交通银行贷款50万，还再还，贷款用于资金周转，接受本机构利息，越快联系越好！
6668	外地人，28周岁，在成都私企工作半年多了，担任质量管理部负责人一职，每月工资收入3000多元。有社保。有使用信用卡，信用好。名下有一套按揭房产（工行，月供1000元，快还完了），83平方，价值50万。贷款用于装修。
5440	长春本地人，客户29岁，名下有全款商品房，房本发票齐全，现单位工作4年，月收入在15000，可提供流水，信用记录良好，在平安贷了5万，还了3个月以上，有意向贷款。



name	gender	year_born	salary	salary_type	company_type	marital	education	record	xd_type	money	month
XXX	男	1985	10000	3.转账	5.私营	已婚	本科	1.良好	1.消费贷	30.00	24

# 数据准备：数据词典

未通过的（主要包含自身条件不符，资料不全，年龄不满，利息太高）（971条）

'不合适','不符','不齐全','高','利率','小额机构不考虑','提供不了','不是很符合','贷不了','不够','不受理','没有办理','不贷款','没有贷款','不通过','不成功','不足','接受不了','不能接受','不接受','被拒','不能提供','拒绝贷款','没有成功申请','没通过','没能受理','没有受理','没受理','未受理','未贷款成功','没贷款成功','未能受理','逾期','没有申请成功','额度','没法贷款','没有批下来','不实','没成功','不能','不想','未满足','问题','没有成功贷款','没有贷款成功','未成功贷款','没成功贷款',



xd\_score=1

xd\_score=1 : 失败  
xd\_score=0 : 成功

# 数据准备: 收集客观数据

## 目前的收集数据

- 1.无信用记录
- 2.信用记录良好
- 3.征信记录较差
- 4.少量逾期
- 5.有信用记录
- 信用记录空

- 1.本地人
- 2.外地人

## 客观的原始数据

- 0.没有逾期
- 1.逾期1次
- 2.逾期2次
- 3.逾期3次
- 4.逾期3次以上
- 逾期空

工作城市

户籍城市

手机所属城市

信贷产品城市

## 衍生的结论数据

- 1. 信用记录好
- 2.信用记录良好
- 3.信用记录较差
- 4.信用记录差

- 1.本地人
- 2.外地人

# 数据准备：数值类数据的收集

月收入
3000-5000
5000-8000

?

4999

5001

月收入
4999
5001

数值型而  
不是区间型

月收入
3千-5千
5千-8千



月收入
4999
5001

数值型而  
不是字符型

数据表 字段	ifre_wide_table 【数据仓库宽表】	ifre_user_info 【用户基本信息表】	ifre_user_comp any_info 【企业主基本信 息表】	ifre_intent_user 【订单表】
性别	√	√	√	
出生年月	√	√	√	
工资	√	√		
公司发放形式	√	√		
企业资产总价值	√		√	
企业月均流水	√		√	
公司类型	√	√ 含(政府/事单位业/国企)	√ 民营/私营/普通	
公司行业	√	√ 含(国家机关)	√	
申请金额	√			√
申请还款时长	√			√
放款时间	√			√



## 数据修复：补性别缺失值

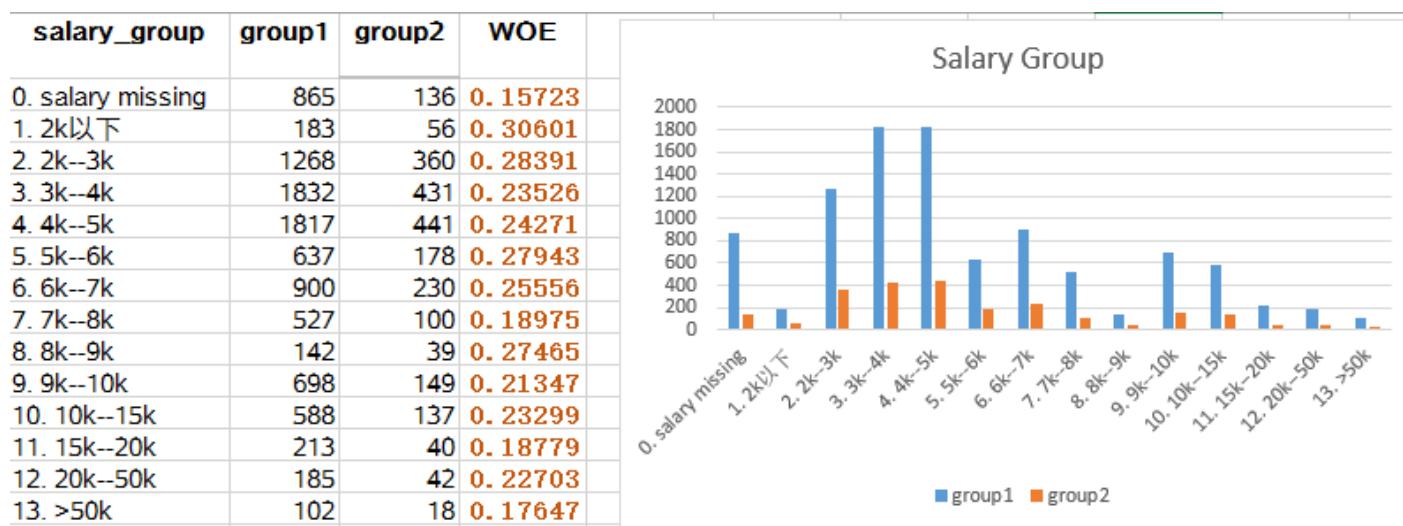
- 目的: 按照real\_name填充缺失性别
- 说明: 按照名字

Eg：先生、女士、易区分性别关键字

含有萍、玲等修复性别为女，含有磊、国等修复性别为男。

状态	男	女	空
初始数量	1718	357	11091
修复后数量	9776	1702	1689

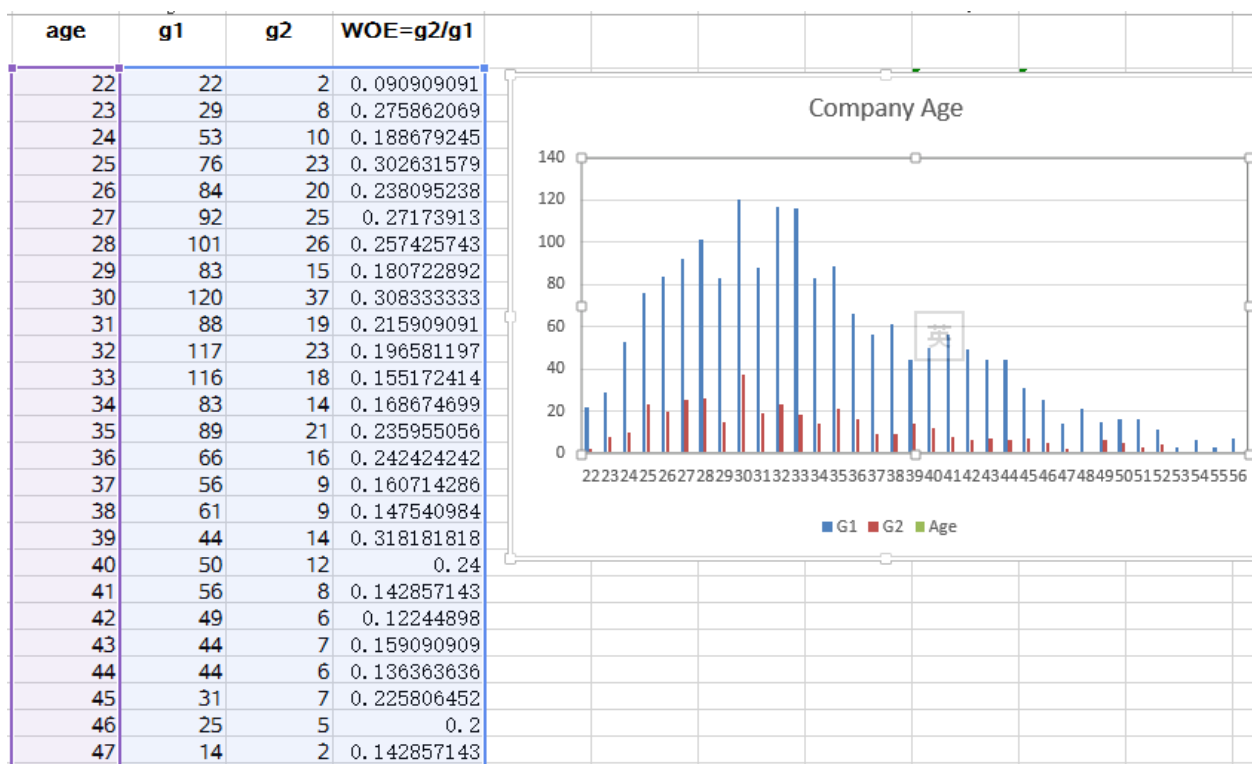
# 数据探索：个人申请-工资 salary



Salary\_group：工资分组  
G1: 贷款失败个数  
G2: 贷款成功个数  
WOE: Weight of Evidence

结论：工资 这个变量并没有对贷款是否成功有显著影响。

# 数据探索：企业主-年龄 age



Age：年龄  
G1: 贷款失败  
G2: 贷款成功  
WOE: Weight of Evidence

结论：年龄 这个变量  
并没有对贷款是否成功  
有显著影响。

# 数据探索：婚姻 marriage

个人申请	marriage	成功	未成功	woe
	1.未婚	183	718	0.25487
	2.已婚	229	902	0.25388
	婚姻空	1945	8335	0.23335
企业主申请	marriage	成功	未成功	woe
	1.未婚	11	102	0.10784
	2.已婚	72	393	0.18321
	婚姻空	461	2109	0.21859

个人和企业主 婚姻 这个变量的缺失值都很高

结论：  
个人的婚姻对成功与否帮助不大；  
企业主已婚的成功率高

# 数据探索：衍生变量

举例：

**( 贷款额度/贷款期限 )**

---

**工资**

# 个人信贷评分卡

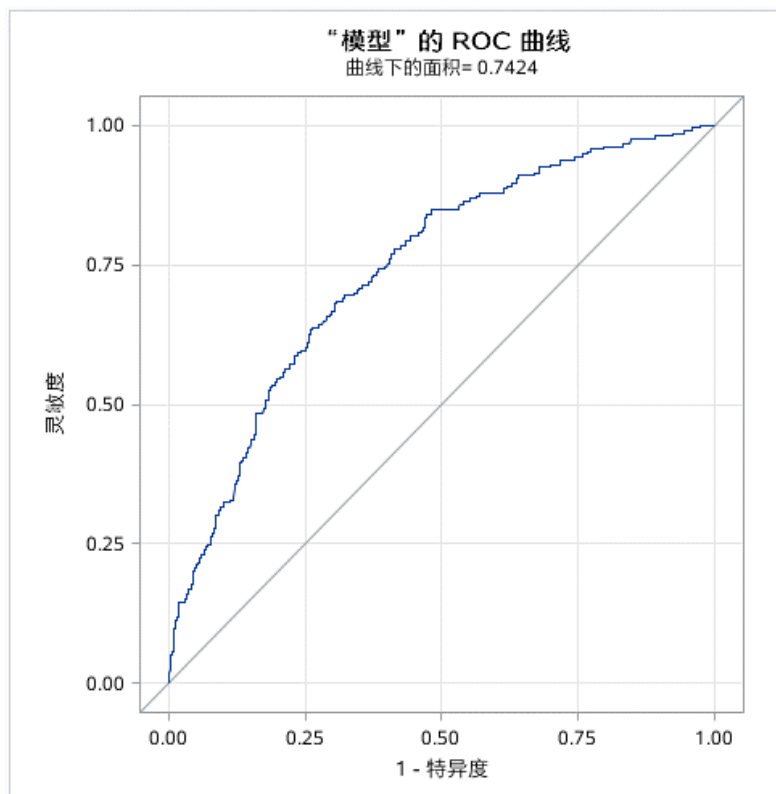
- 算法：Fisher Linear Discriminant，决策树，逻辑回归 Logistic Regression，神经网络
- Logistic回归是研究因变量为二分类或多分类观察结果与影响因素（自变量）之间关系的一种多变量分析方法，属概率型非线性回归。
- $\ln(P/(1-P)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots$  Odds =  $P/(1-P)$

$$P = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_m X_m)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_m X_m)}$$

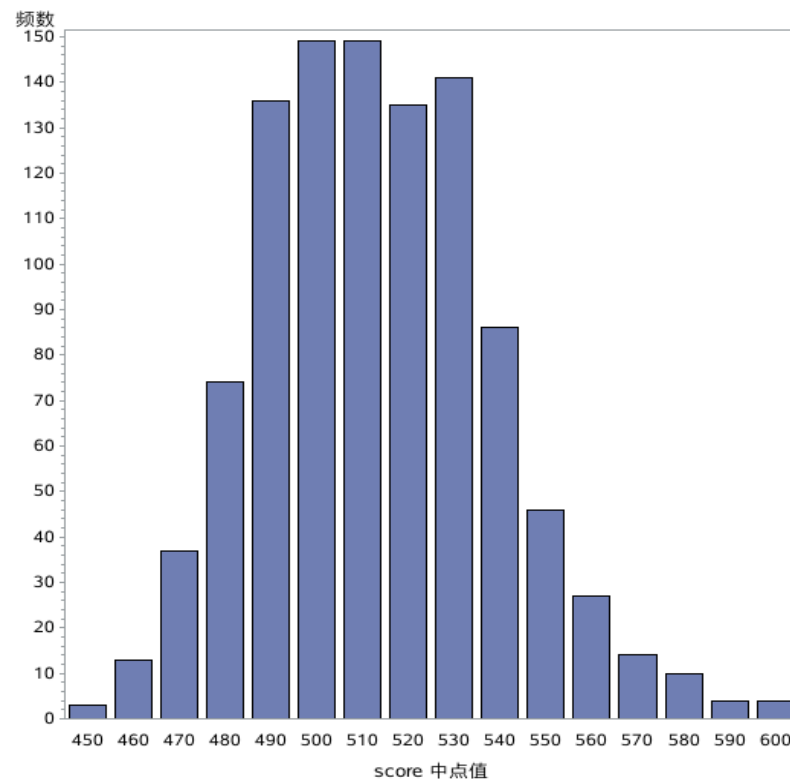
其中， $\beta_0$  为常数项， $\beta_1, \beta_2, \dots, \beta_m$  为偏回归系数。

# 个人信贷评分卡

个人信贷评分卡



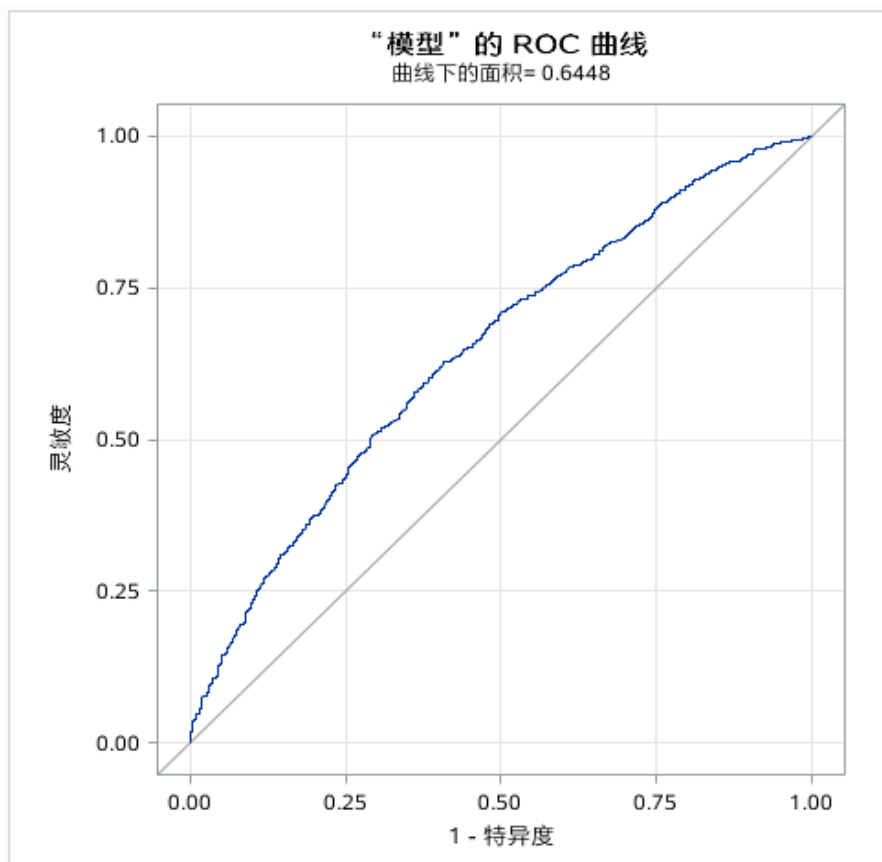
个人非业主模型ROC曲线



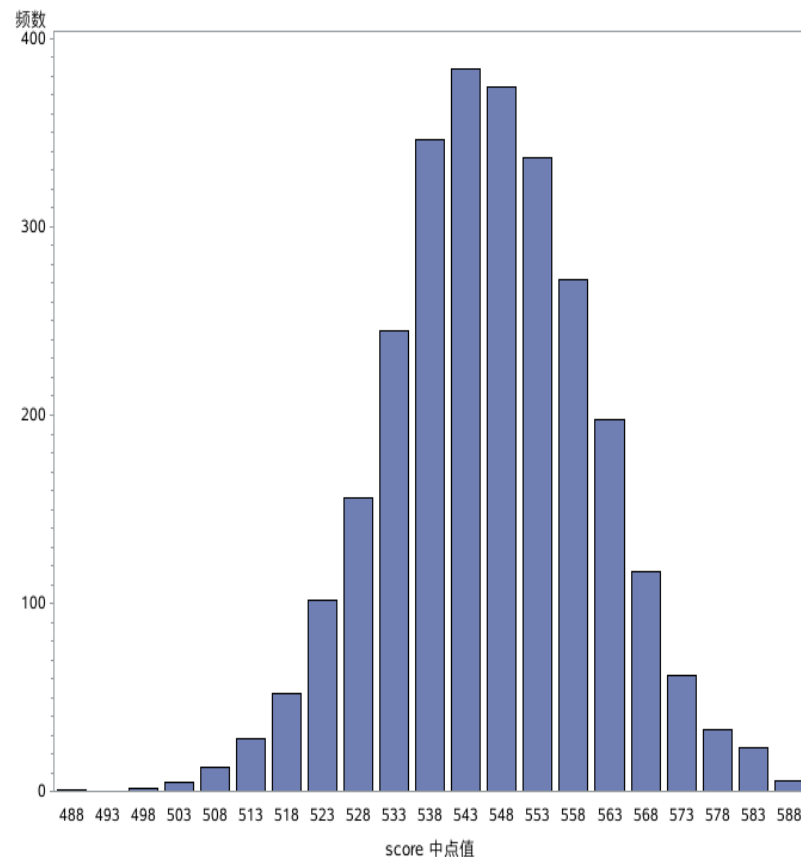
个人非业主模型分数分布图

# 企业主信贷评分卡

- ◆ 分数范围：[480, 600]
- ◆ 稍左偏



企业主模型ROC曲线

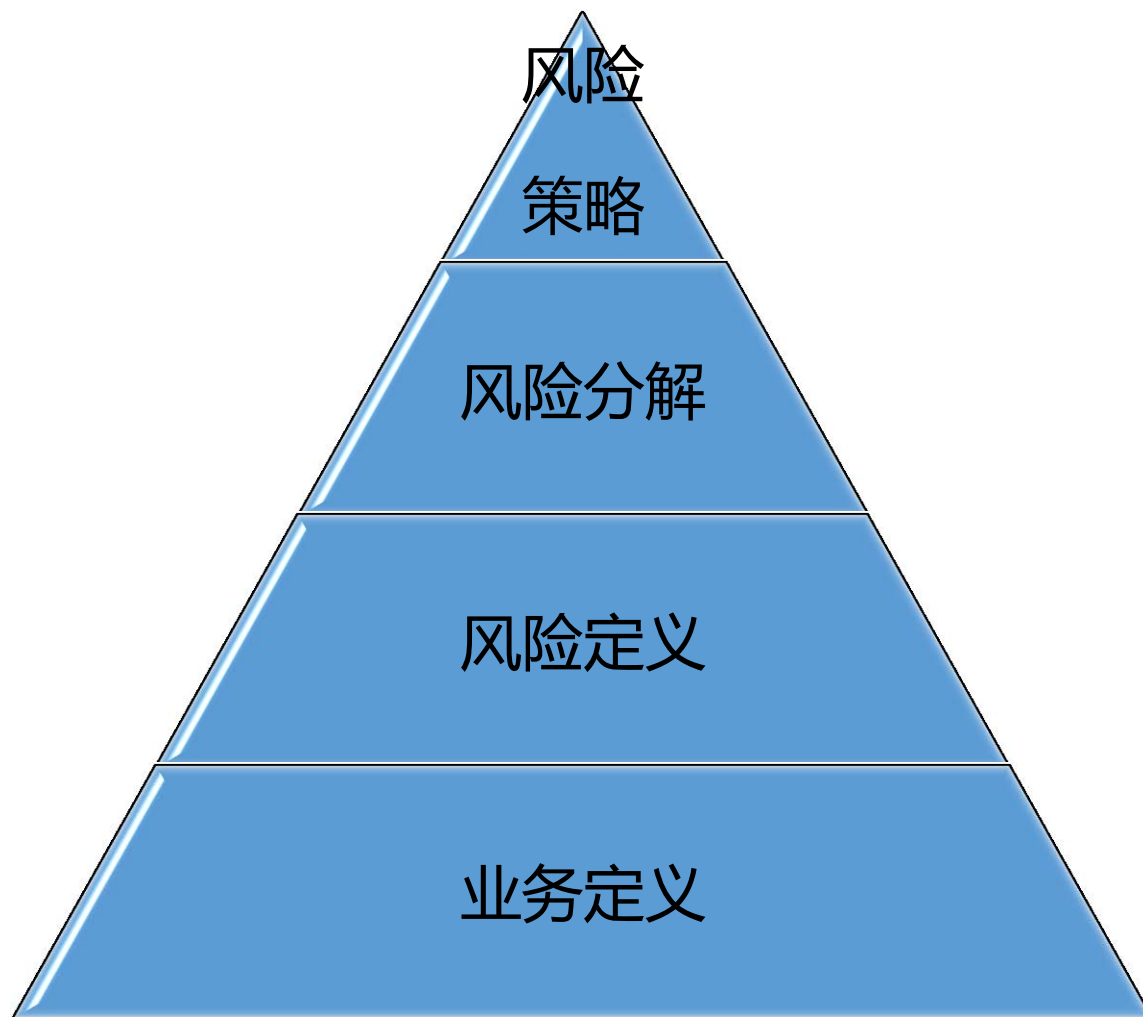


企业主模型分数分布图

# 模型验证

模型	K-S
A. 授薪产品模型（个人非业主）	34.6%
B. 业主产品模型（个人业主）	27.6%
C. 经营产品模型（企业主）	19.4%

# 模型应用策略



# 模型应用策略

- 产品定价：按照模型计算的通过率给申请信息定价
- 流程简化：通过模型分类，降低审核人员的工作量
- 精准营销：通过分析信贷员的准过率，同样类型的信息优先推荐给放款审核宽松的信贷员
- 商业模式拓展：对通过率高的客户，可直接推荐给B端，放款收佣金
- 上述策略的应用基础建立在模型精准的基础上，现有的模型需要新的数据迭代验证和迭代开发。

量化资产

定价风险

促进流动

-----数信互融



评分系统微信账号

公司公众账号



IFRE