

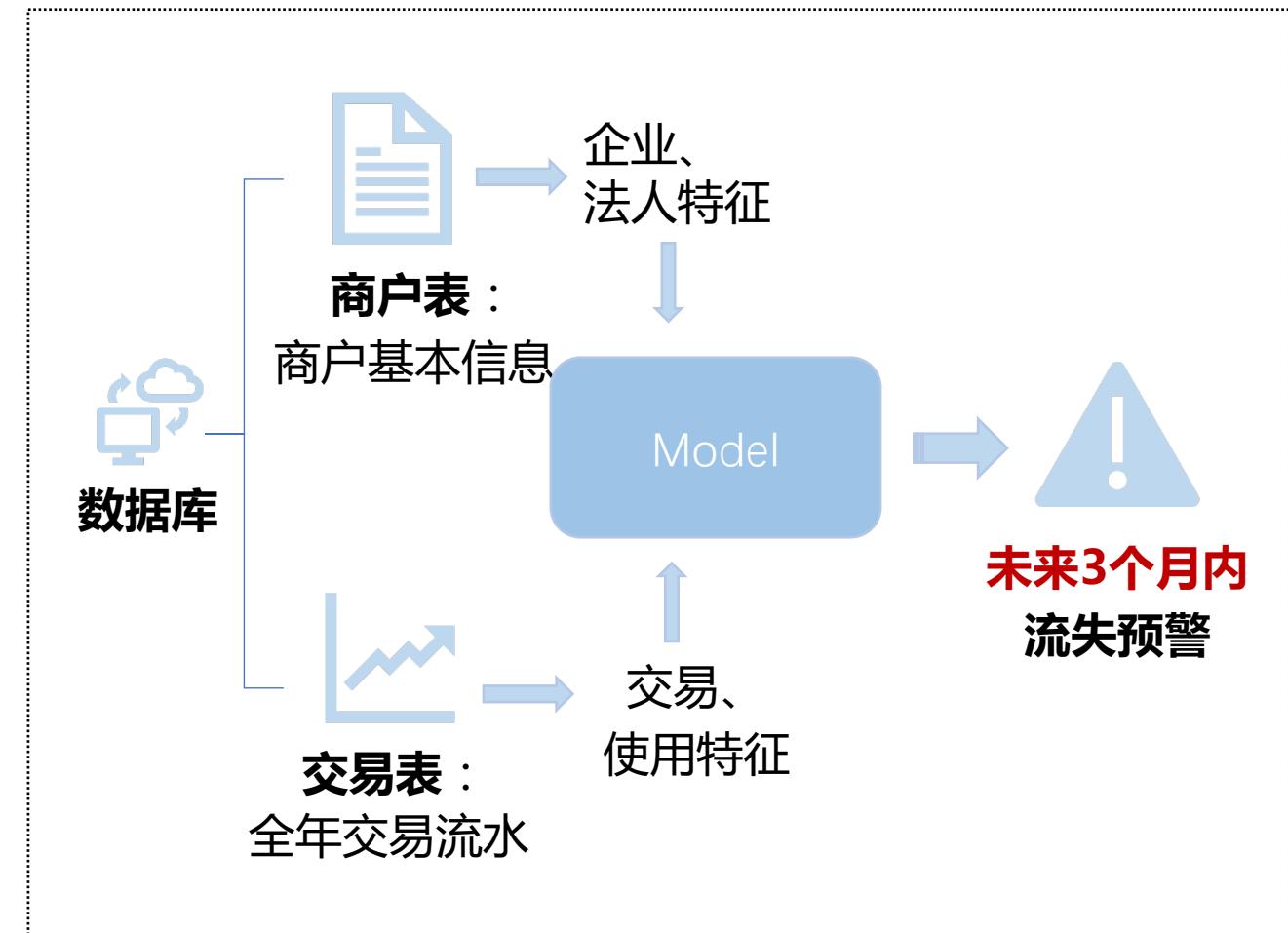
# 流失预警模型

——不活跃商户分析及预测

□ 数据规模：

注册商户数量：**100000**

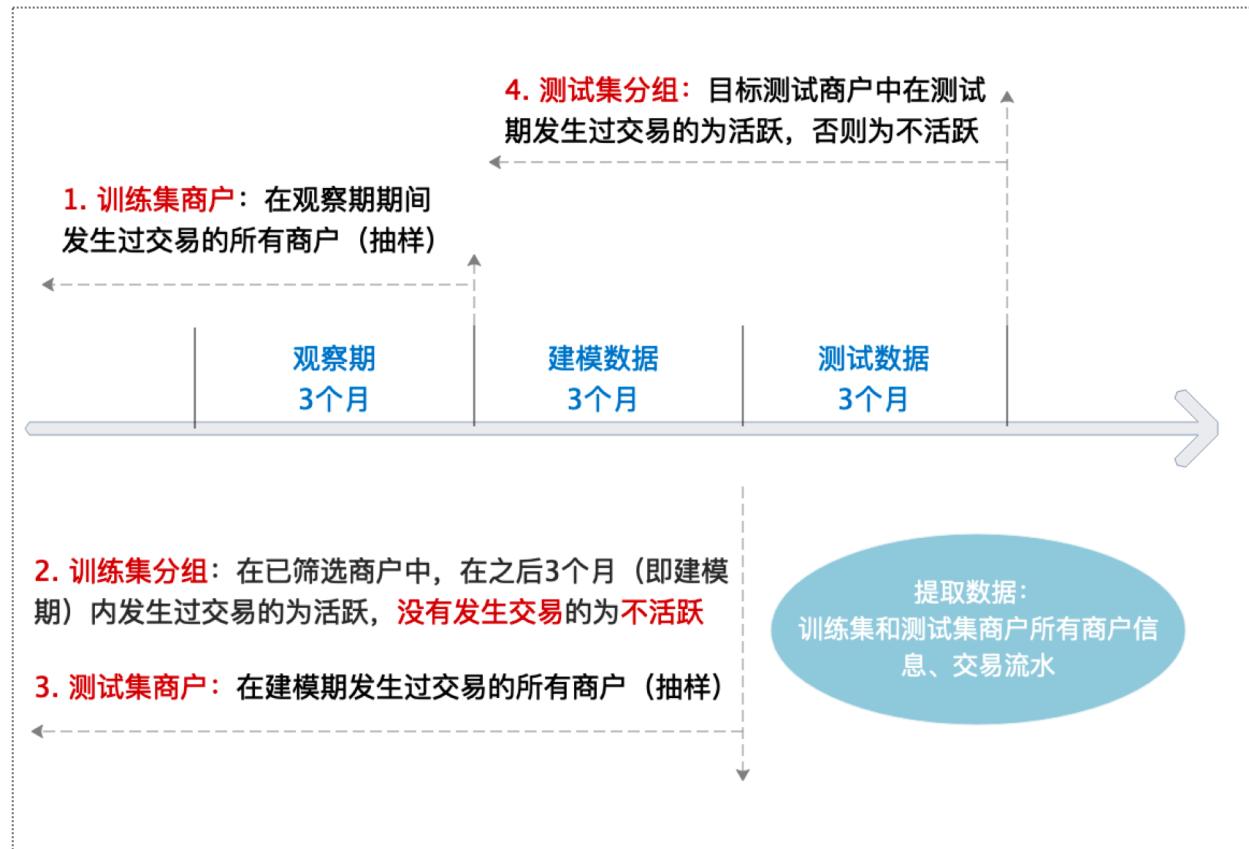
18年全年交易流水：**5亿+**



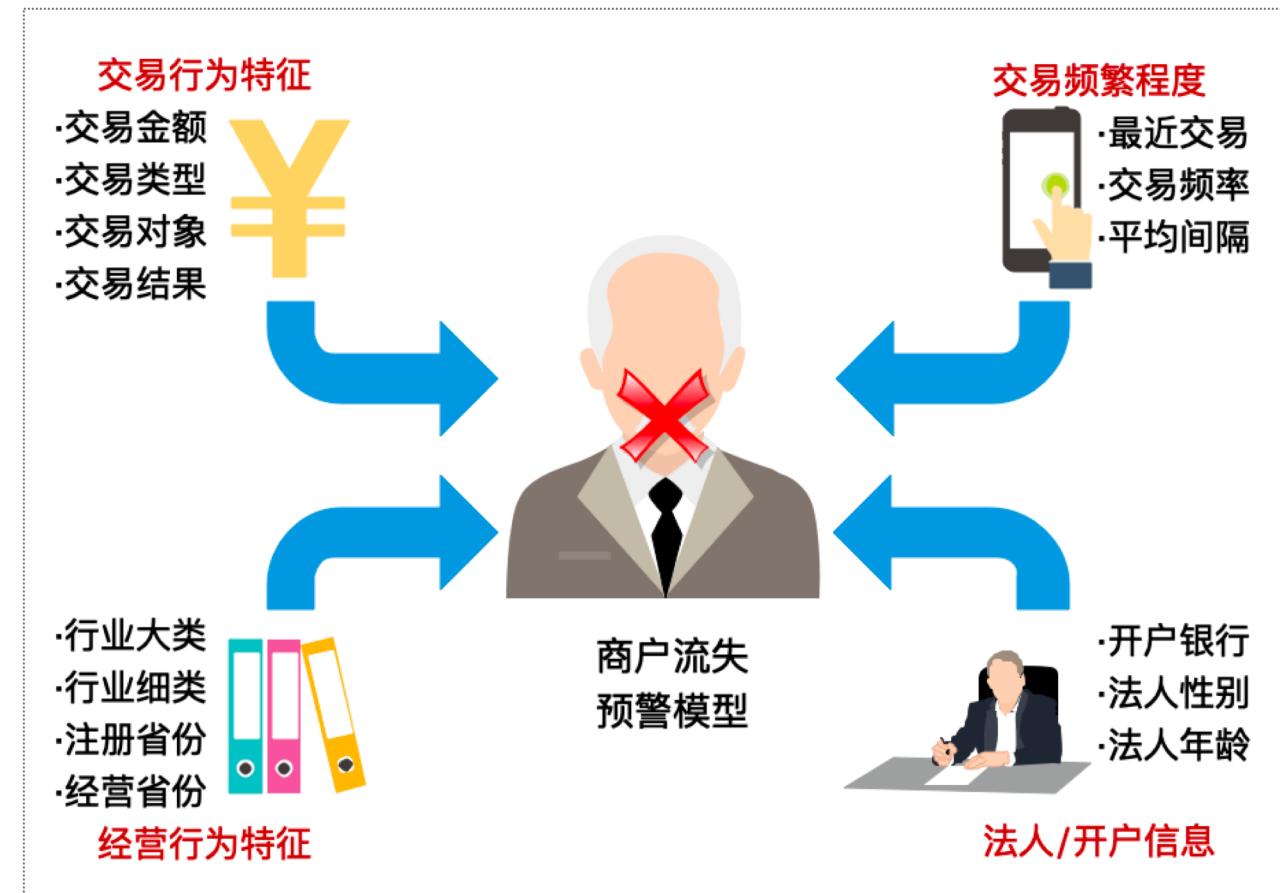
□ 目标商户：在观察期的三个月内发生过交易，但可能会在未来三个月流失的潜在商户。

□ 抽样方法：

- 训练集：不活跃商户和活跃商户分别等量抽样，得到10000个不活跃样本及10000个活跃样本。
- 测试集：在建模期发生过交易的商户中随机抽取10%，得到10000个测试样本。



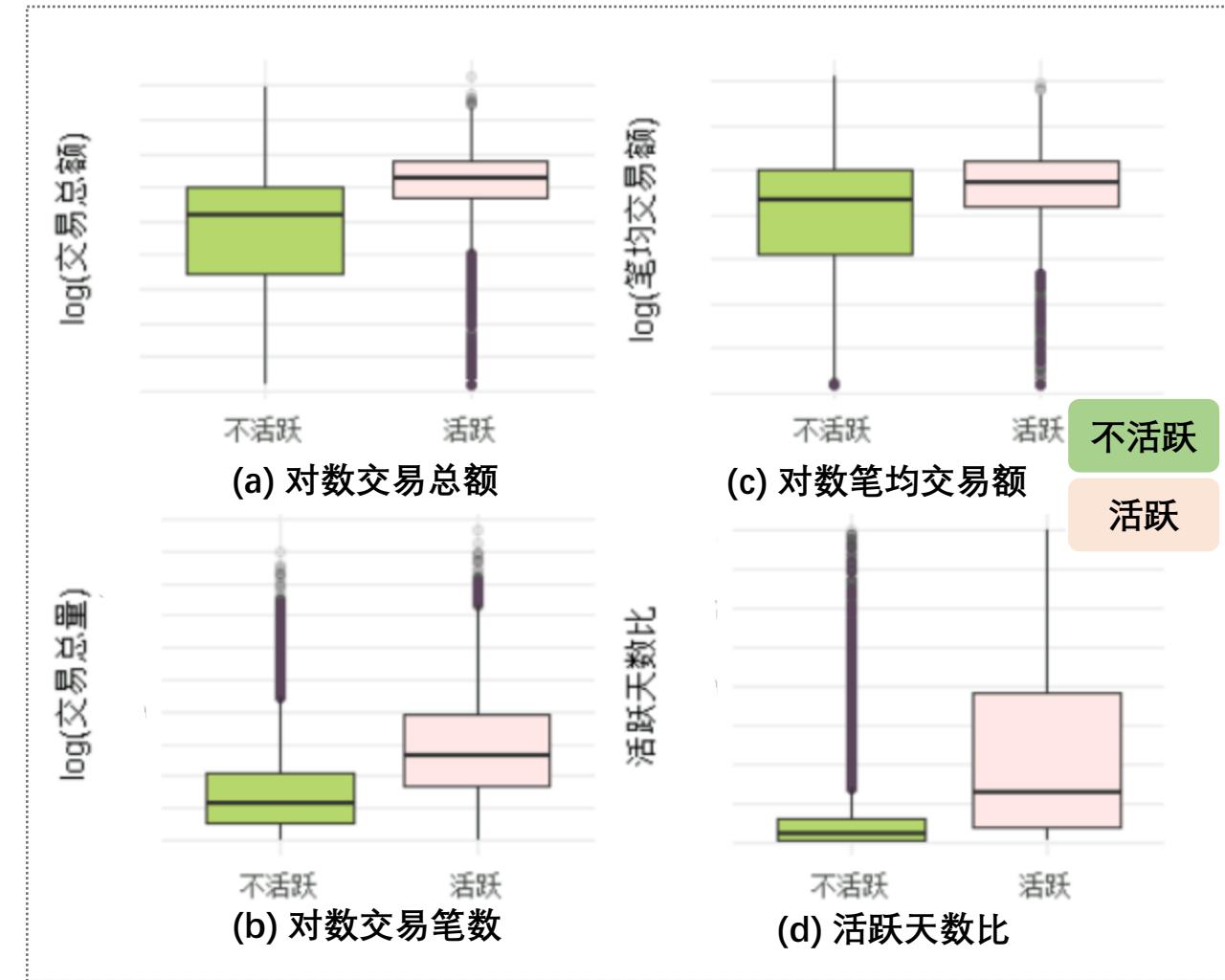
- 静态指标：商户的注册信息，商户的经营信息，法人的个人信息，行业信息...  
从商户表中提取、分类。
  
- 动态指标：交易信息、登陆信息...  
从交易流水中计算新的指标，如交易额、交易量、活跃天数等。



变量类型	变量名	详细说明	取值范围	备注
因变量	是否不活跃	定性变量 共2个水平	1代表不活跃 0代表活跃	训练集1:1
经营特征	行业细类	定性变量 共10个水平	餐饮类等	训练样本少于 100归为“其他”
	经营省份	定性变量 共23个水平	上海市等	训练样本少于 100归为“其他”
自变量	交易总额	单位：元	0-10000000	
	交易总量	单位：笔	0-100000	
	笔均交易额	单位：元/笔	0-1000000	
	活跃天数比	单位：无	0-1	

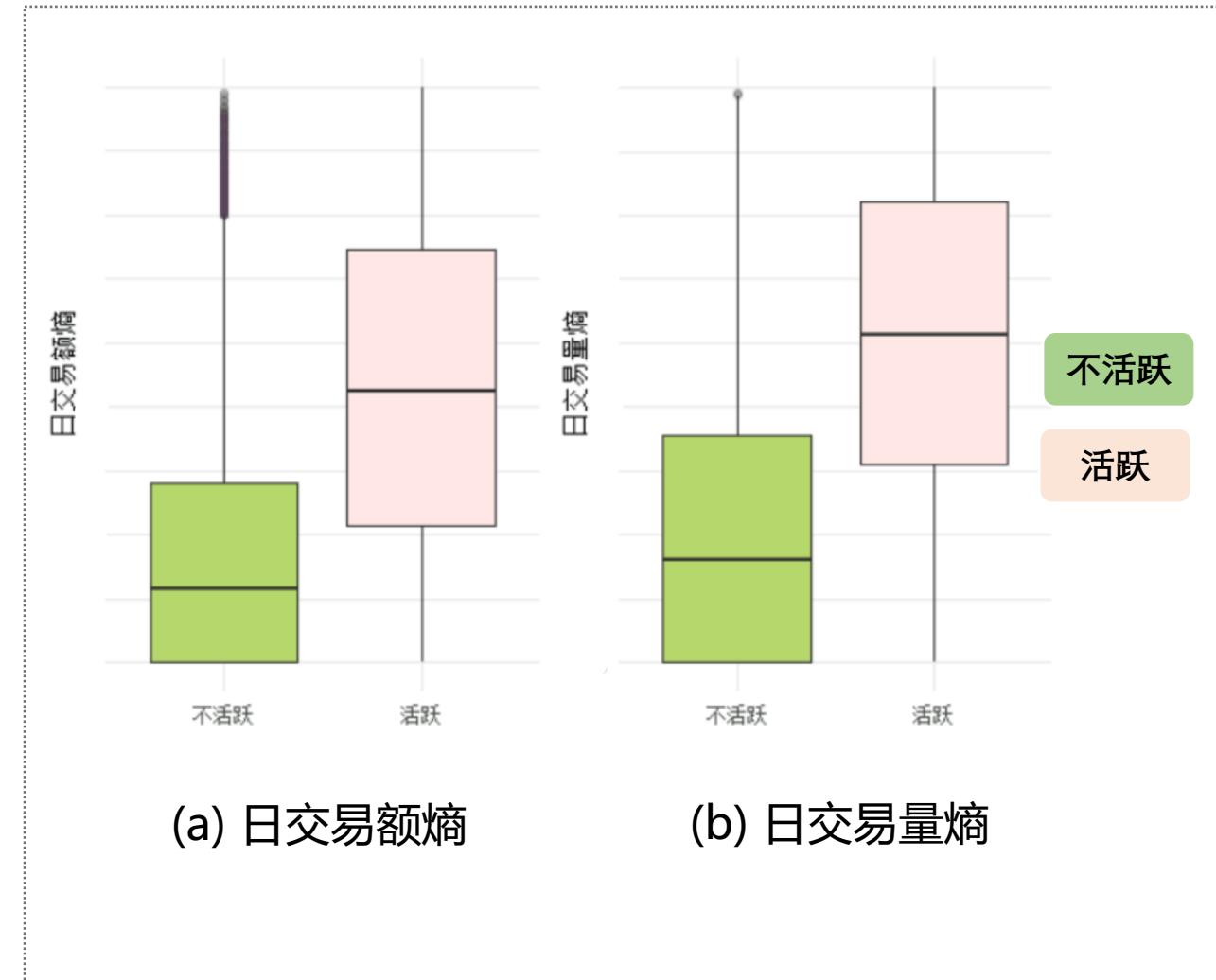
变量类型	变量名	详细说明	取值范围	备注
自变量 交易行为特征	日交易额熵	单位：无	0.00-5.00	
	日交易量熵	单位：无	0.00-5.00	
	凌晨交易量占比	单位：无	0-1	0:00 ~ 6:00
	上午交易量占比	单位：无	0-1	6:00 ~ 12:00
	下午交易量占比	单位：无	0-1	12:00 ~ 18:00
	晚间交易量占比	单位：无	0-1	18:00 ~ 24:00

- 不活跃商户在**交易总额、交易笔数、笔均交易额、活跃天数比**上都明显低于活跃商户。

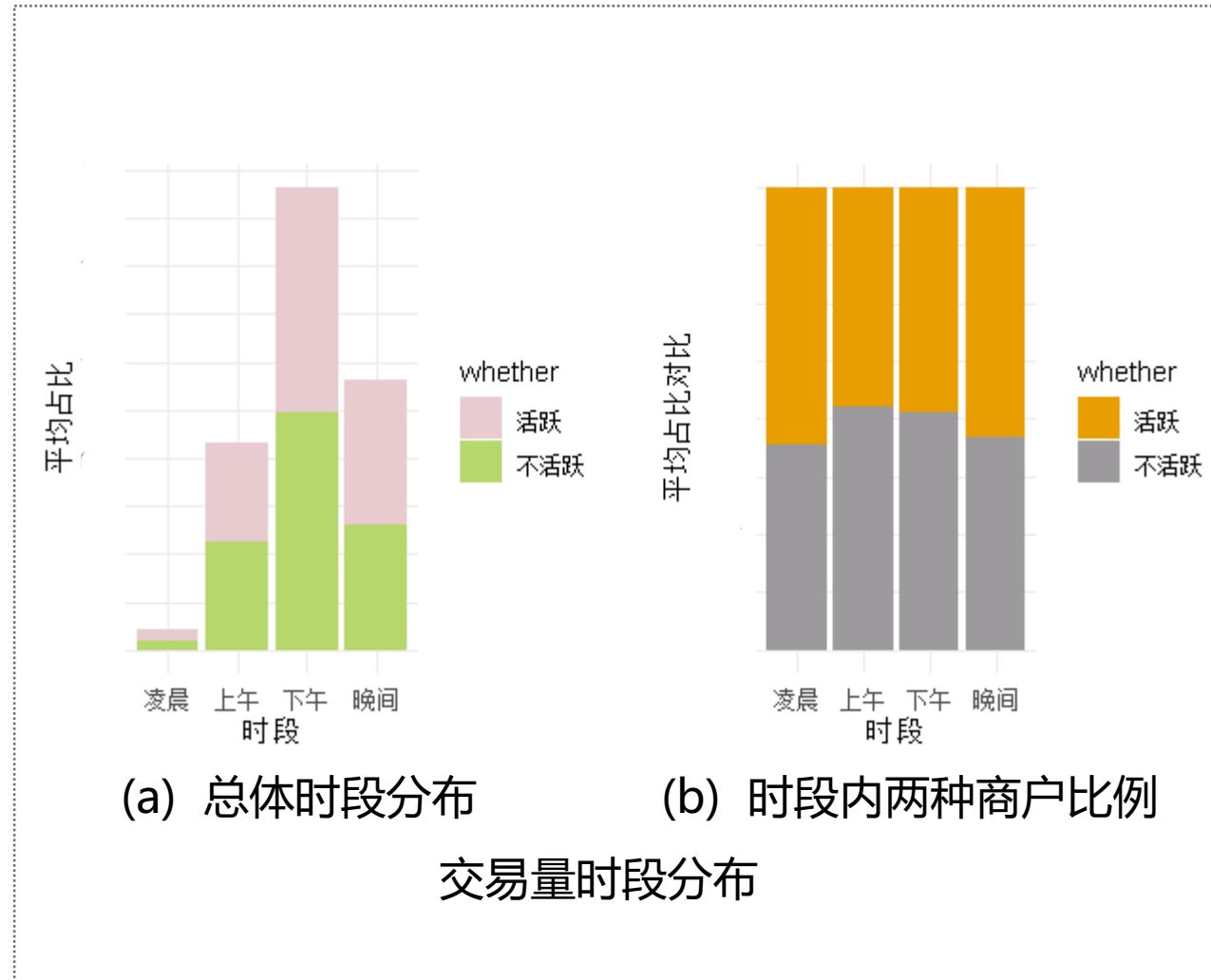


□ 不活跃商户日交易额熵、日交易量熵明显较低；

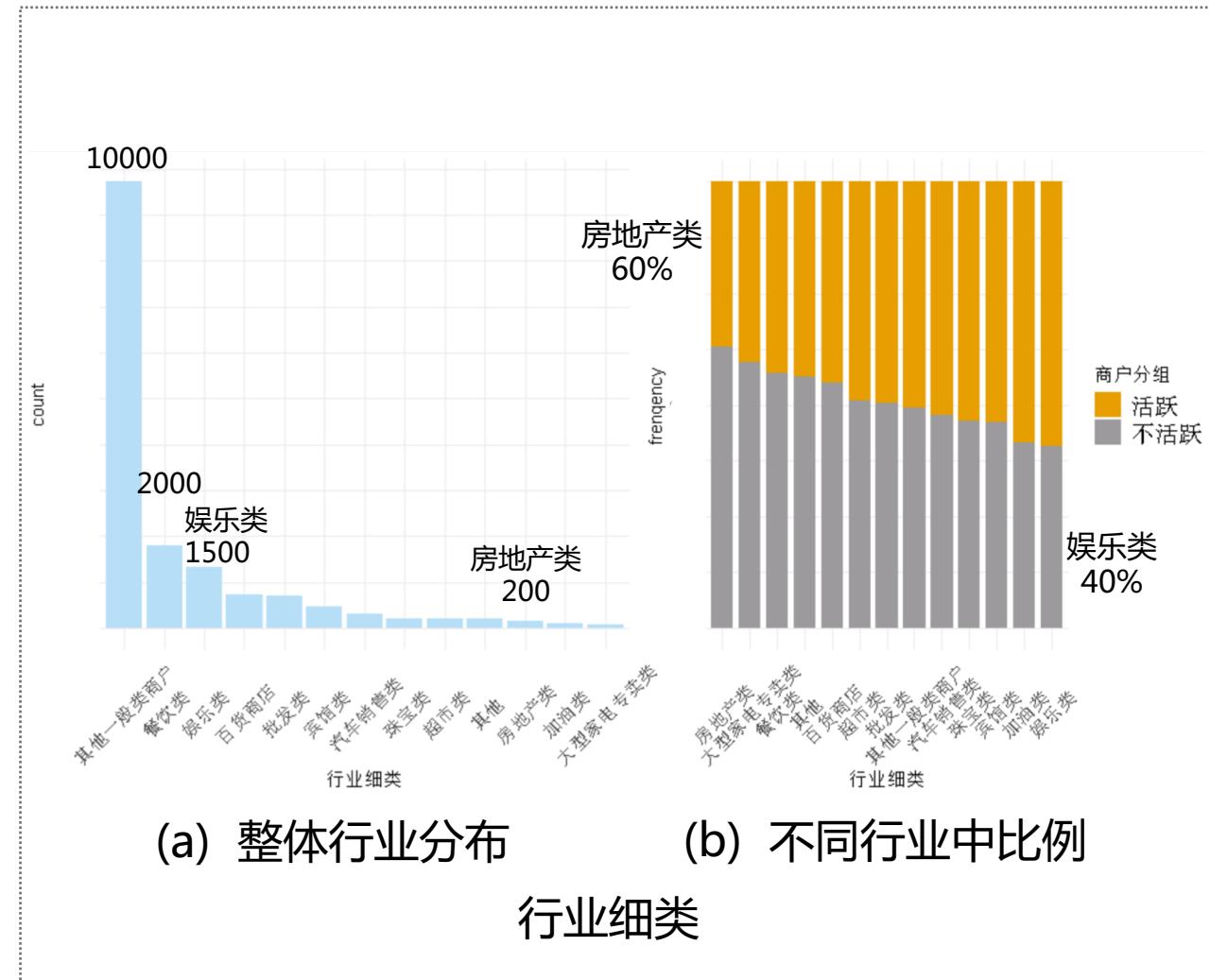
□ 熵值越高说明交易发生的日期越分散，即稳定性越好。



- 大部分的交易量发生在下午时段，凌晨时段交易量很小；
- 不活跃商户的交易量发生在上午和下午的比例更高。



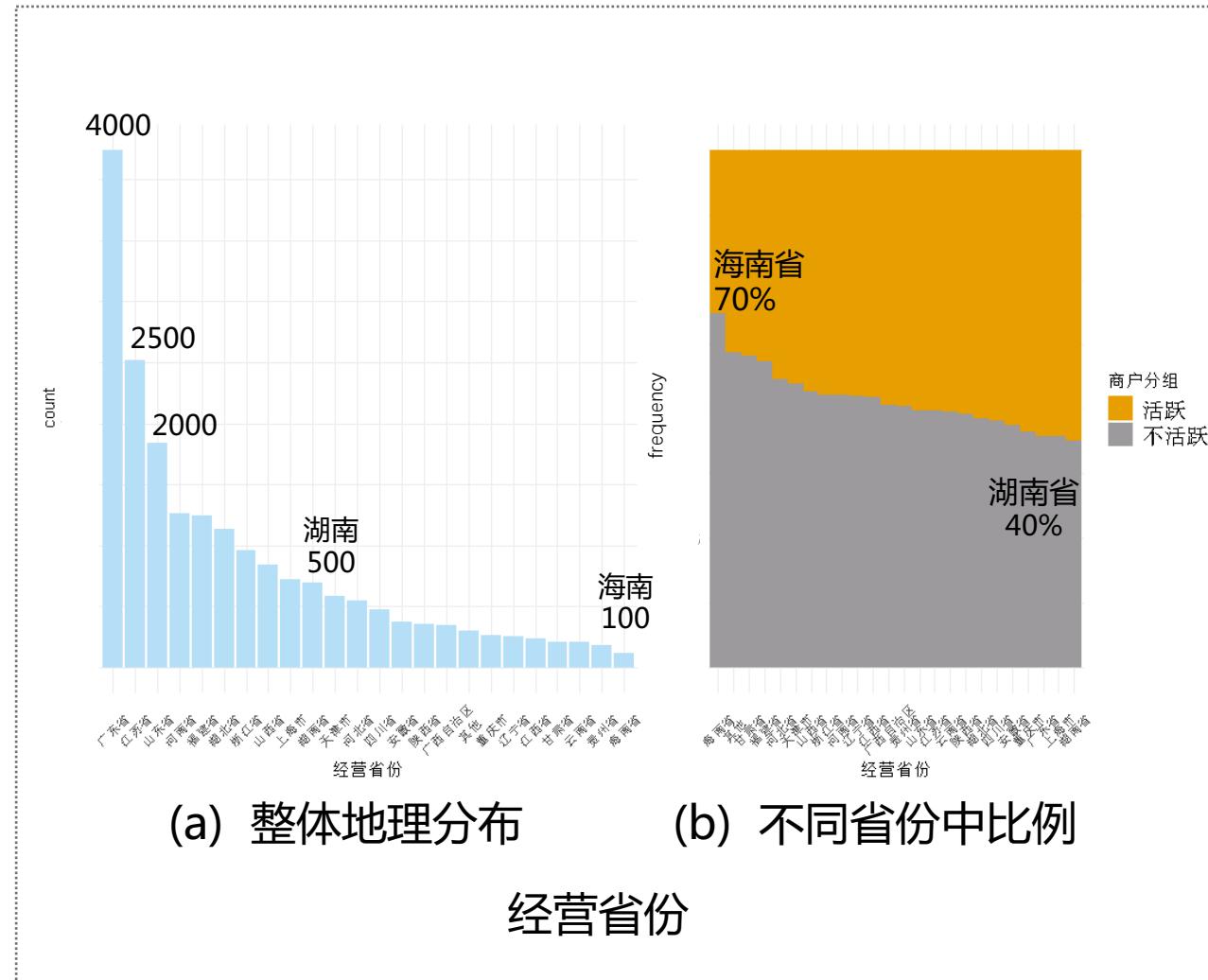
- 从事**其他一般类**行业的商户明显**多于**其他行业
- **房地产类**行业的**不活跃**比例**较高**，**娱乐类**和**珠宝类**行业的不活跃比例**较低**。



□ 广东省的经营商户**最多**，其次是江苏和山东；

□ 海南省的商户**不活跃**比例**远远高于**其他省份，

湖南省的商户**不活跃**比例**最低**；



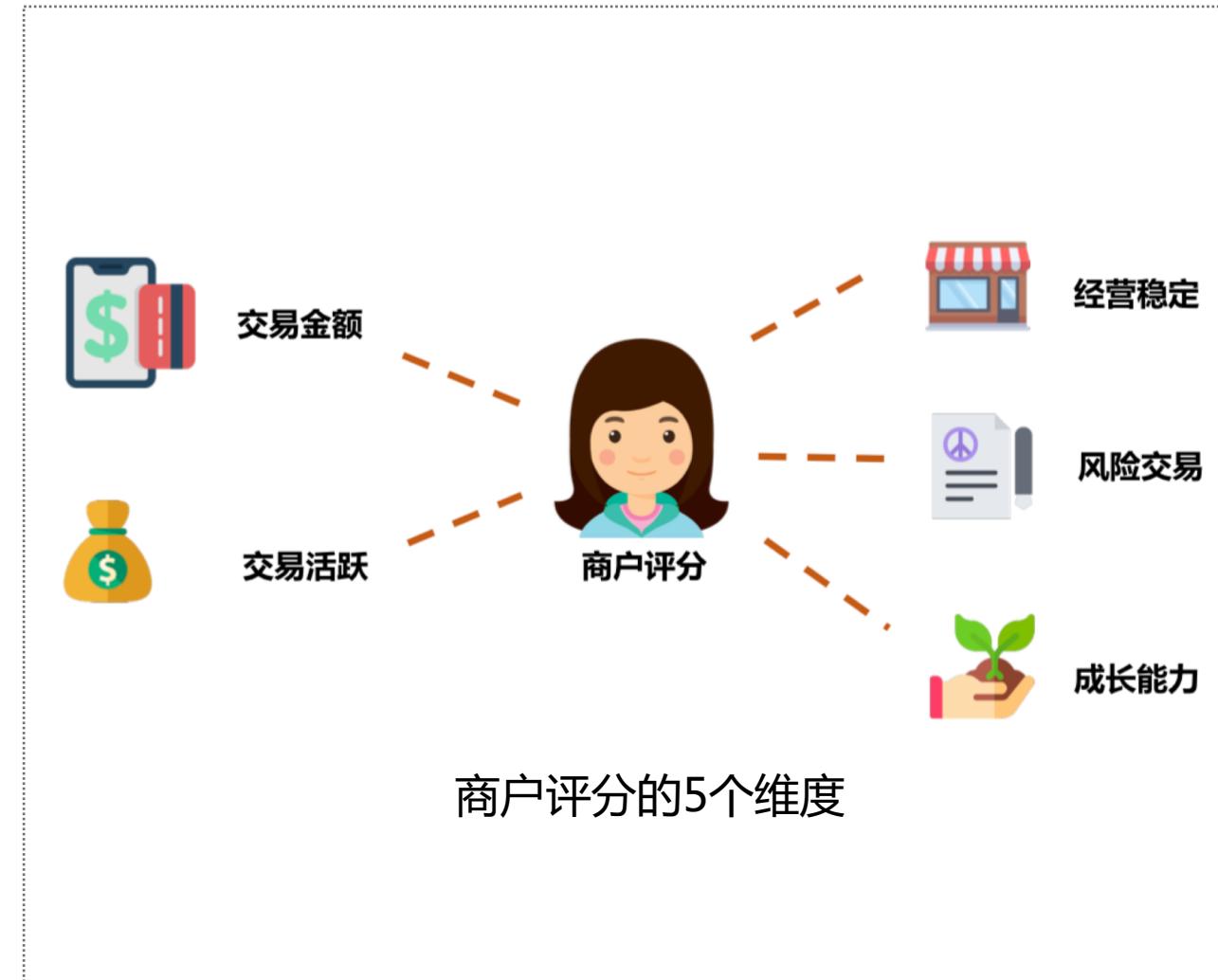
# 进阶指标

风险打分

多维度打分

评价商户风险水平

- 每个维度中包含1个或多个指标，若其中至少一个指标得分为0，则这个维度总评分为0；
- 每个维度内的指标按**AHP层次分析法**计算权重。



- 对于数值越大，风险越低的指标，采用**分位数打分**，如交易金额；
- 对于数值越大，风险越大的指标，采用**反向分位数打分**，如最近交易间隔时间；
- 对于数值偏离均值越多，风险越大的，采用**中值打分**，如离散系数。

分值取值为 0-100，**分值越高，商户风险指数越低**

对每个商户基于**同行业比较**进行打分，共有3种打分方法：



#### 分位数打分

- 取值高于99%的用户，则记为99分
- 适用于大部分指标



#### 反向分位数打分

- 按从小到大排序，如果排名高于99%的用户，则记为99分
- 针对波峰日交易金额/平均日交易金额等指标



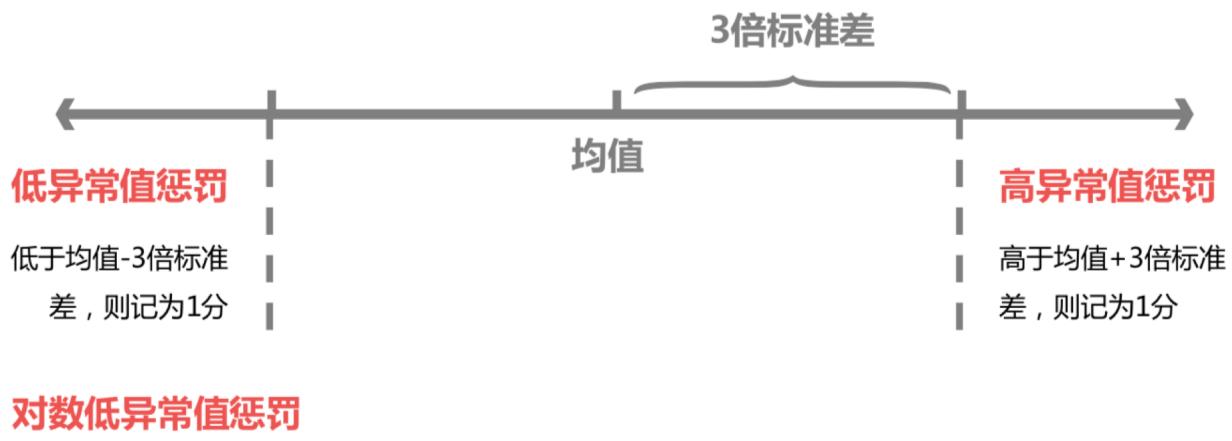
#### 中值打分

- 计算公式： $100 - \text{abs}(\text{取值} - \text{均值}) * 100 / (\text{标准差} * 3)$
- 取值等于均值时为100分；差距越大则打分越小，负数取1
- 针对离散系数等指标

每个指标打分的3种标准

- 低异常值惩罚：活跃天数等
- 高异常值惩罚：离散系数等
- 对数低异常值惩罚：  
交易金额等跨度特别大的指标

另有3种惩罚方式：



对于极端异常值的惩罚方式

## 商户评分-交易金额得分

TECHNOLOGY CO., LTD.

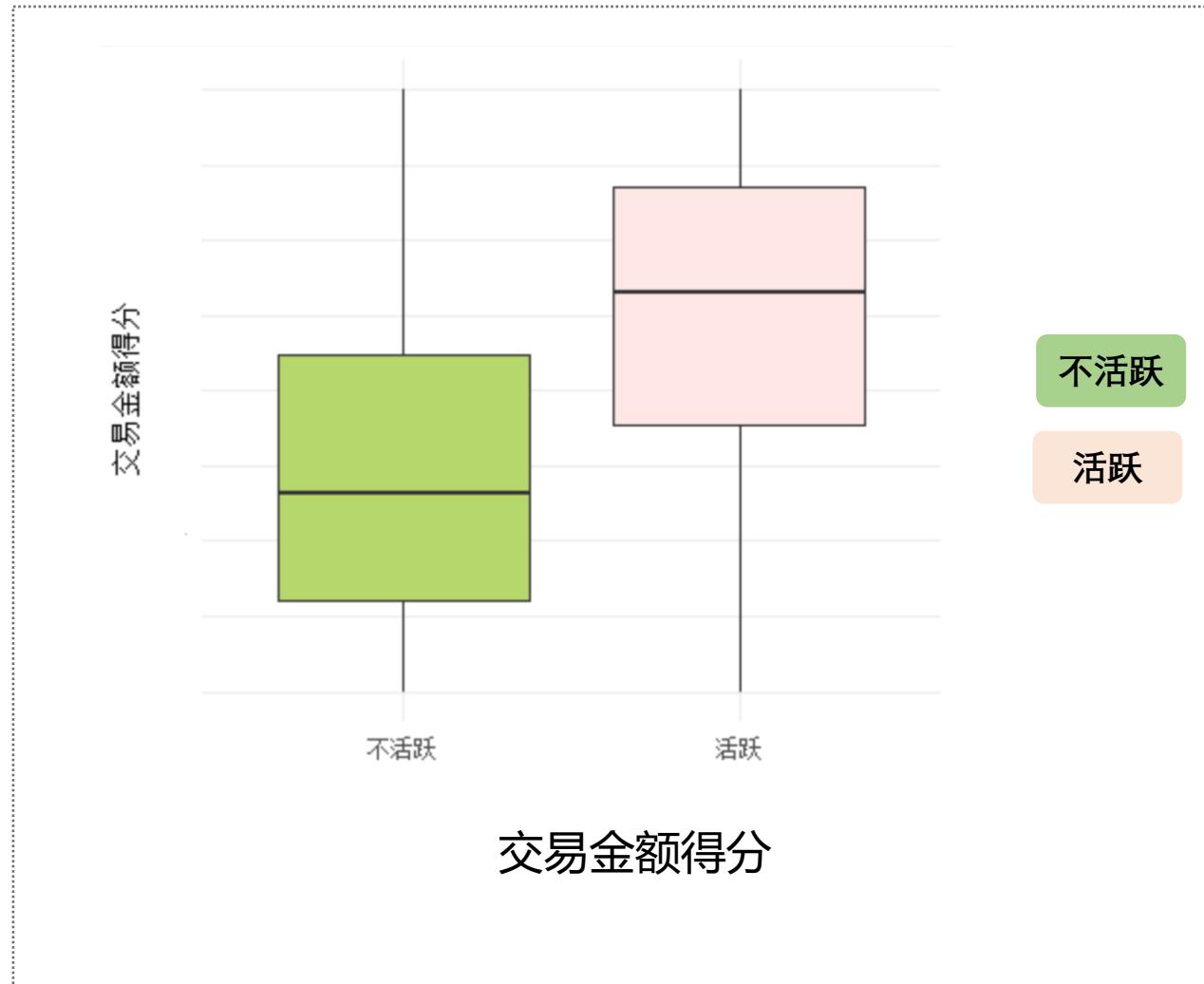
□ 分位数打分，对数低异常值惩罚；

□ 包含指标：

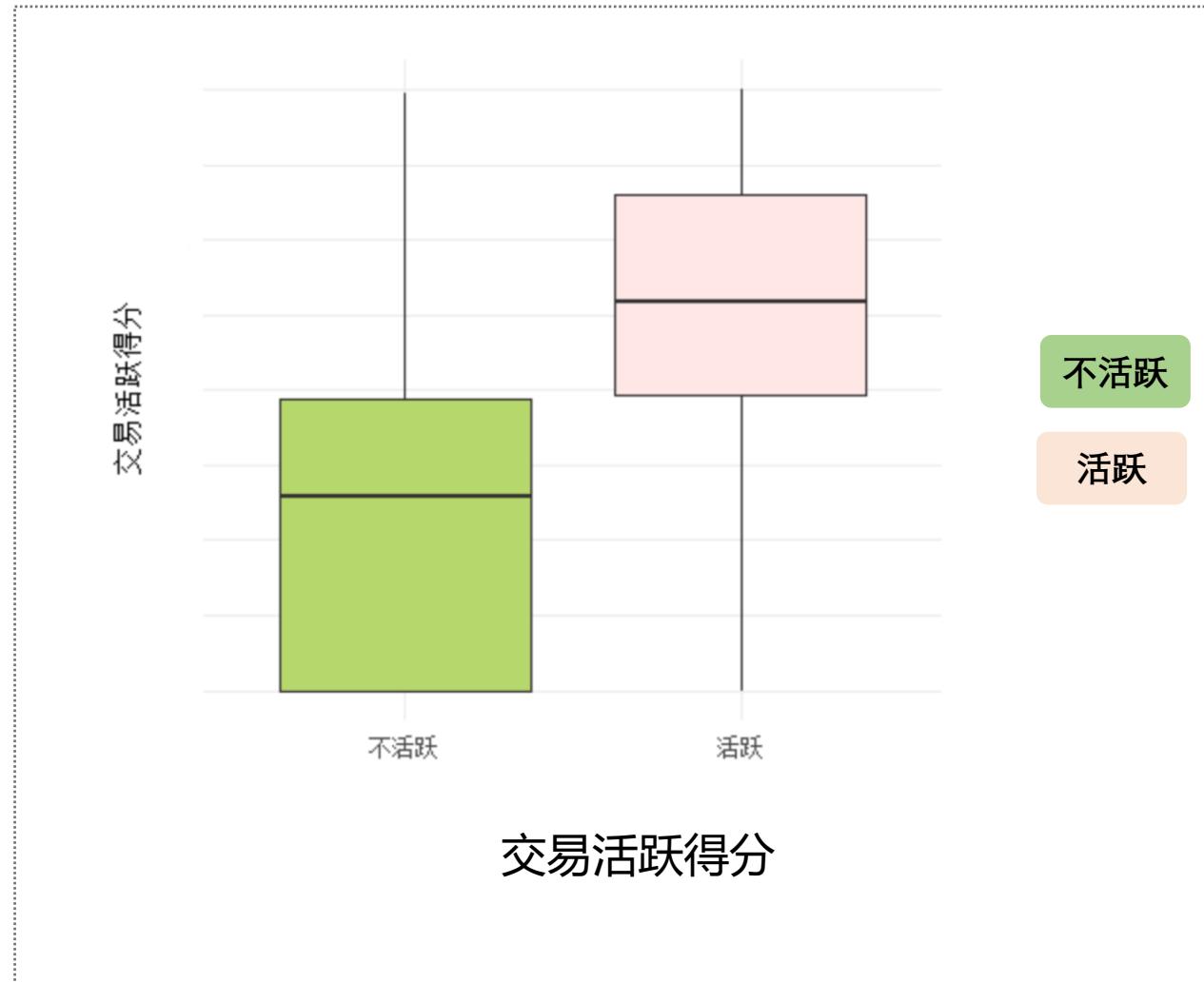
( 1 ) 3个月内交易金额；

( 2 ) 活跃天数内日均交易额。

□ 不活跃商户得分明显**偏低**。



- 分位数打分，低异常值惩罚；
- 包含指标：
  - ( 1 ) 最近一笔交易时间间隔；( 反向分位数 )
  - ( 2 ) 3个月内活跃天数；
  - ( 3 ) 活跃天数内日均交易量。
- 不活跃商户得分明显**偏低**。



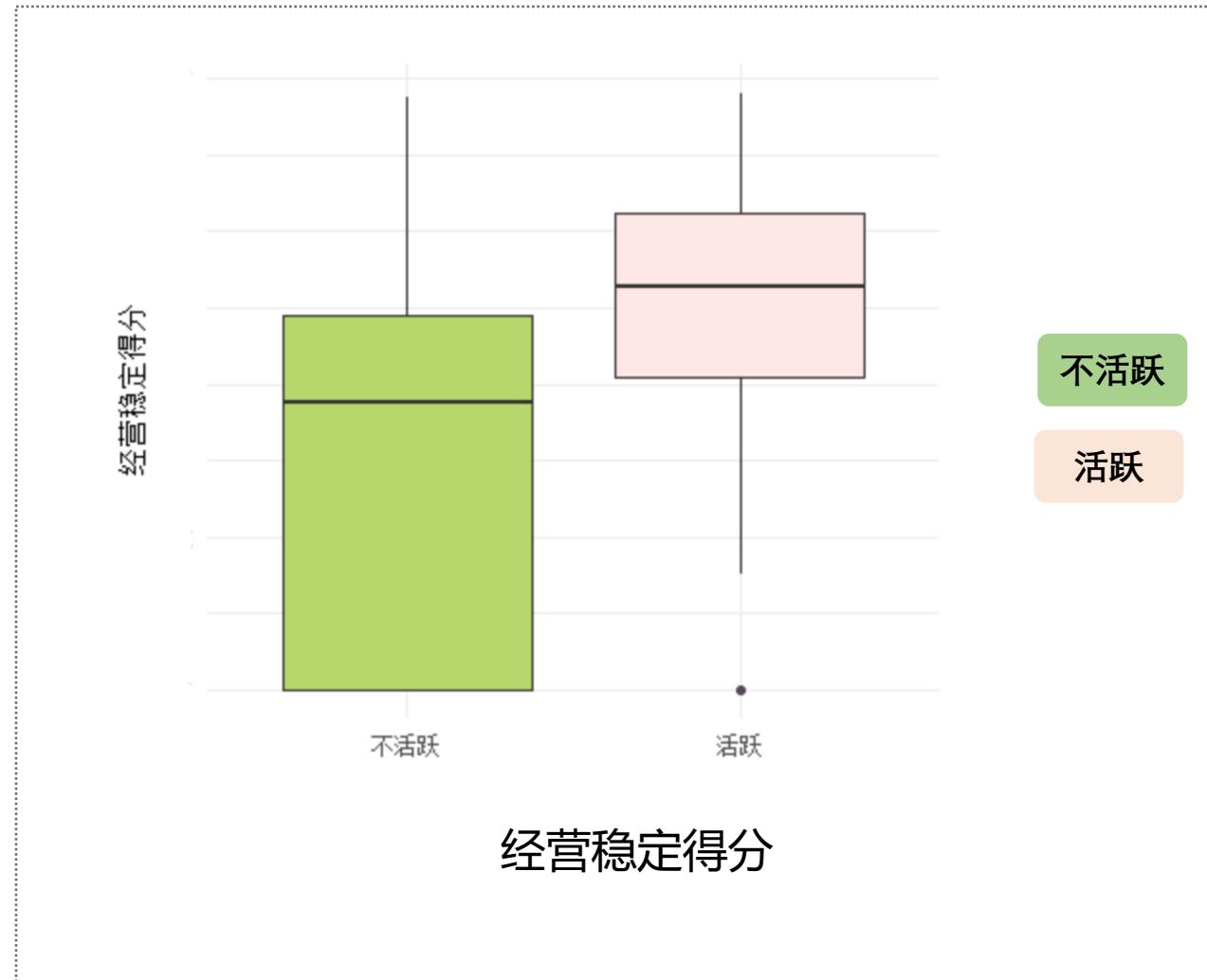
## 商户评分-经营稳定得分

TECHNOLOGY CO., LTD.

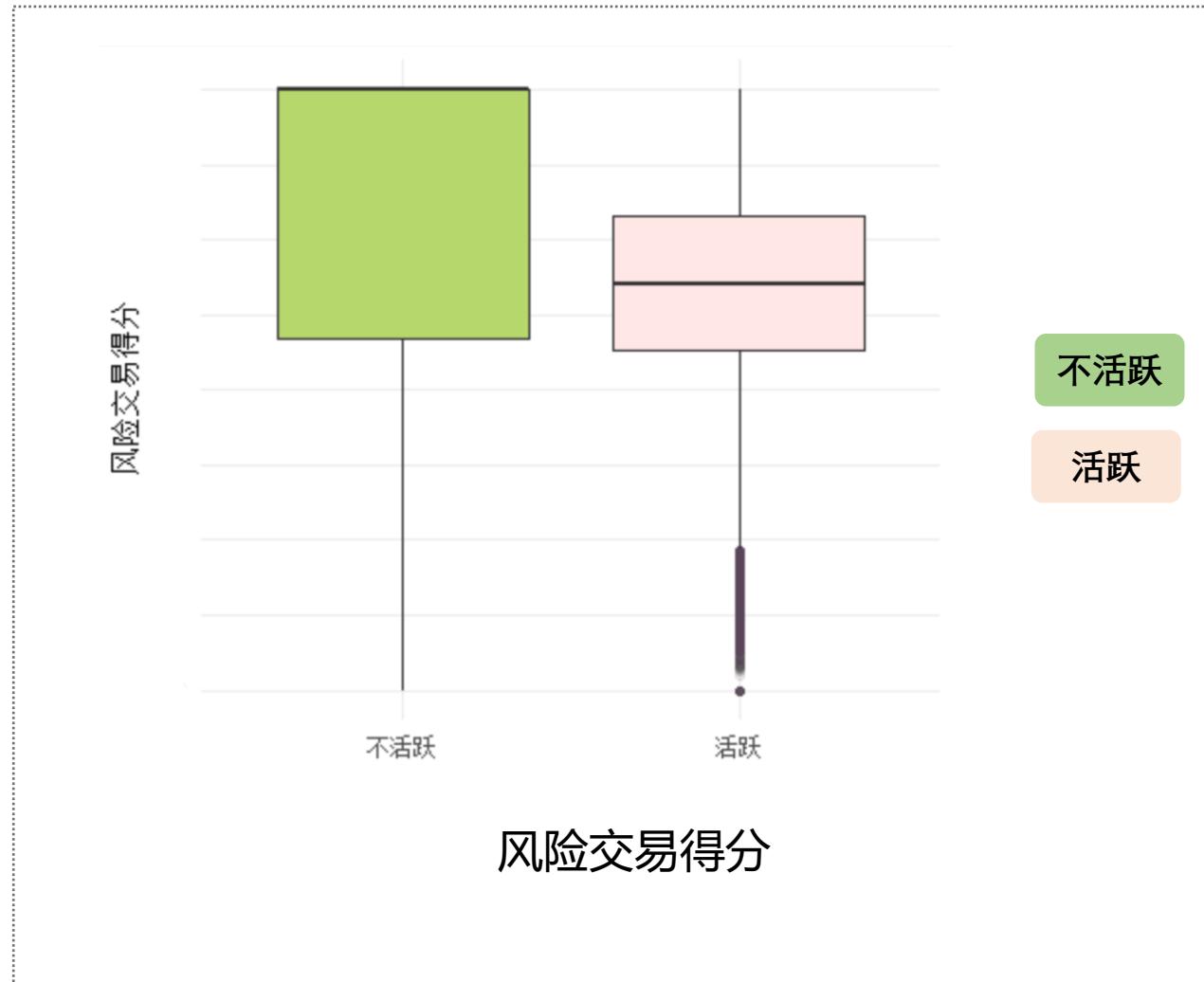
□ 包含指标：

- ( 1 ) 日交易额熵；
- ( 2 ) 日交易量熵；
- ( 3 ) 入网时间；
- ( 4 ) 离散系数。 → 中位数打分，  
分位数打分，  
低异常值惩罚  
高、低异常值惩罚

□ 不活跃商户得分偏低，得分分散。



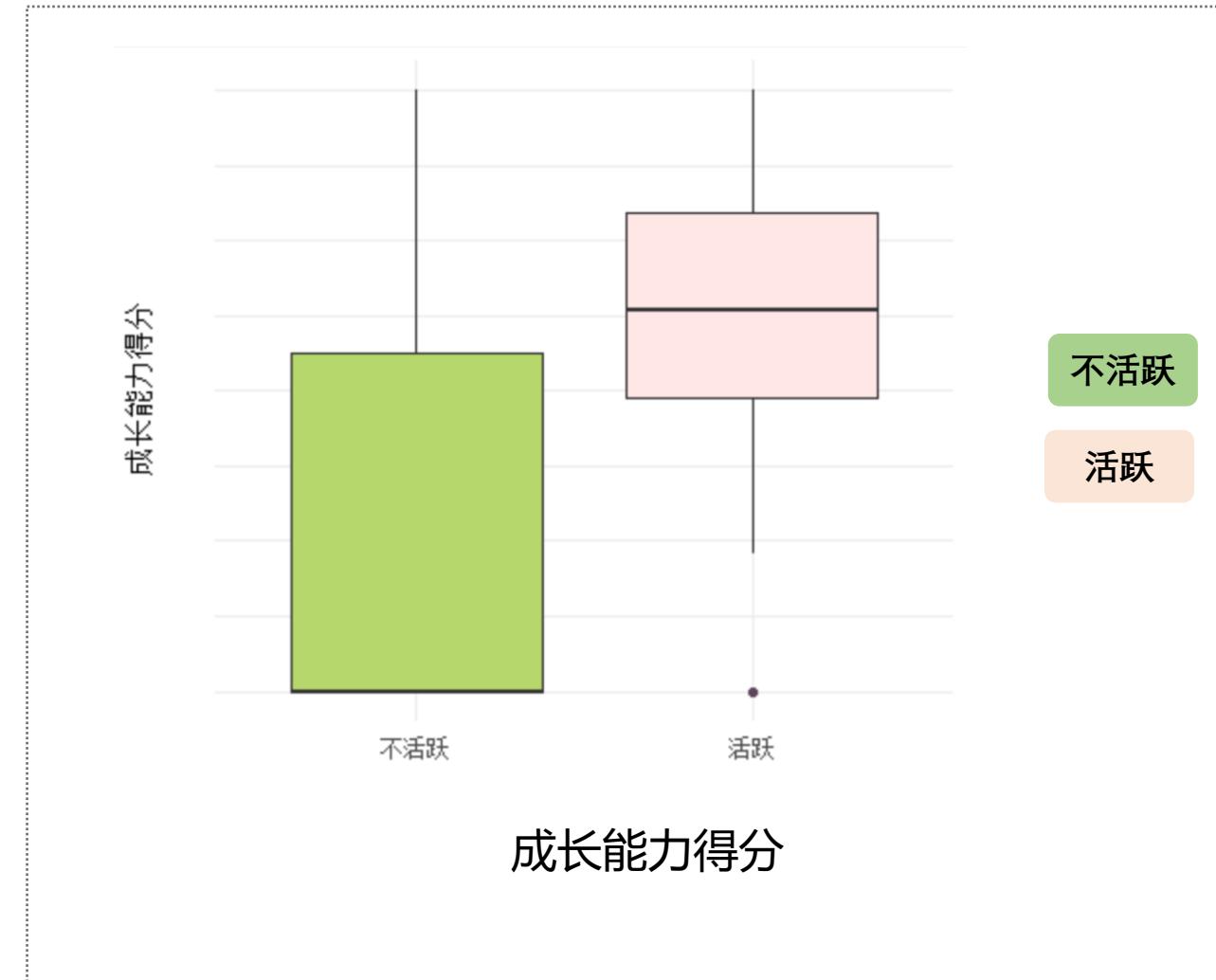
- 反向分位数打分；
- 包含指标：
  - ( 1 ) 失败交易占比
  - ( 2 ) 非工作时间交易额占比。 ( 0:00-6:00 )
- 不活跃商户得分**很高**。
- 不活跃商户发生交易数少，发生风险交易概率更小。



## 商户评分-成长能力得分

TECHNOLOGY CO., LTD.

- 正向分位数打分，低异常值惩罚；
- 包含指标：  
( 1 ) 近1月内交易金额 / 近3月内交易金额。
- 不活跃商户得分普遍**很低**。



# 建模分析

逻辑回归

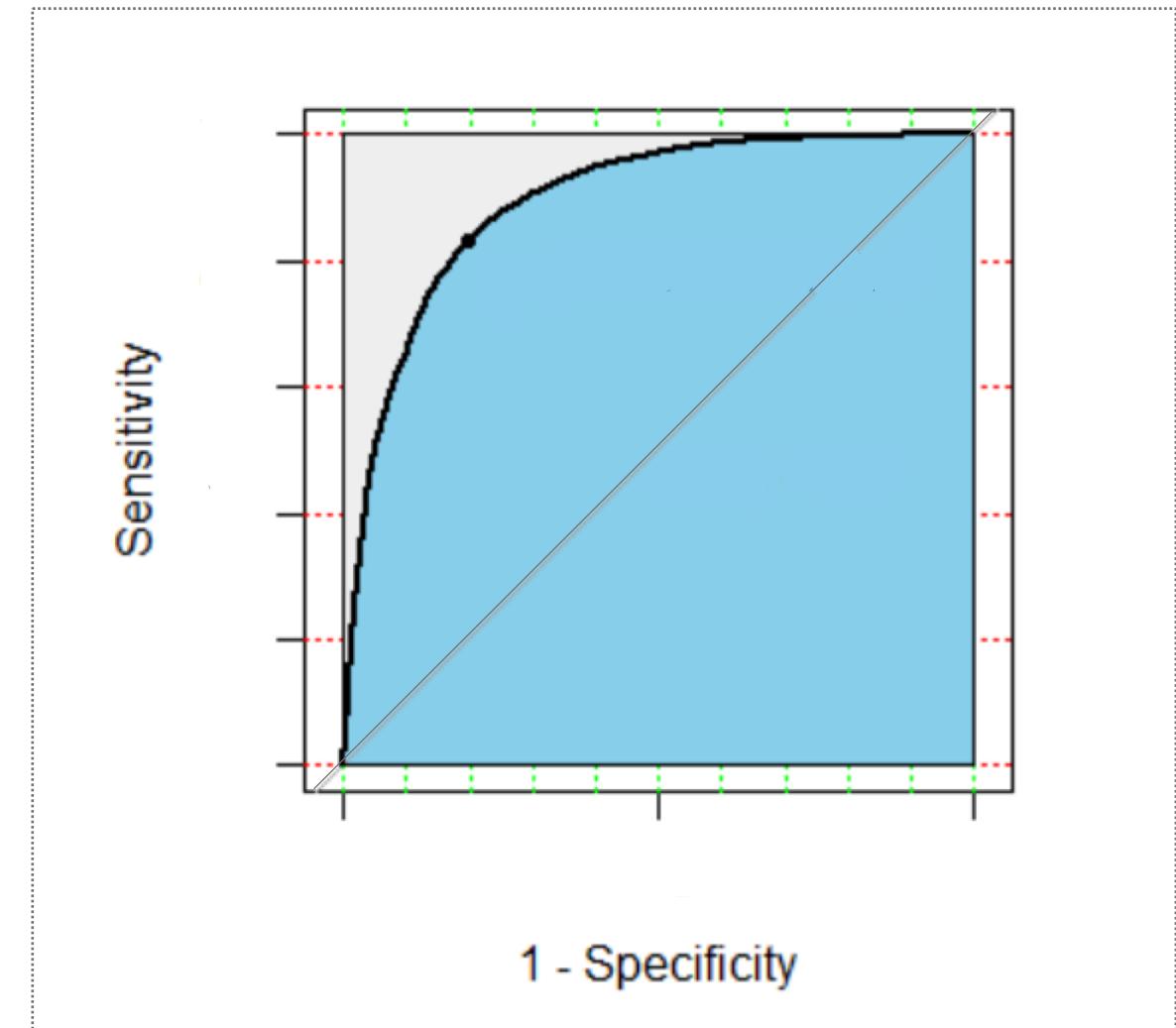
随机划分80%为训练集

20%为测试集

交易行为特征	变量名称	回归系数	P-value	显著程度	备注
交易指标	(截距项)	1	<0.001	***	
	Log笔均交易额	-1	<0.001	***	
	log交易总量	1	<0.001	***	
	活跃天数	-1	<0.001	***	
交易量占比	凌晨	1	<0.001	***	
	下午	-1	0.01	*	
	日交易额熵	-1	<0.001	***	
	日交易量熵	-1	<0.001	***	
商户评分	交易金额得分	-1	<0.001	***	
	交易活跃得分	-1	<0.001	***	
	经营稳定得分	1	<0.001	***	
	成长能力得分	-1	<0.001	***	

		变量名称	回归系数	P-value	显著程度	备注
经营特征	经营省份	广东省	-1	0.01	*	基准：其他省
		海南省	1	<0.001	***	
		江苏省	-1	0.1	.	
		山东省	-1	1		
	行业	餐饮类	1	<0.001	***	基准：其他类
		娱乐类	-1	<0.001	***	
		珠宝类	1	1		

- 模型所用的变量根据AIC准则选出。
- 凌晨和下午交易量占比、熵和商户评分均显著；加入这些变量后， $\log$ 交易总量的影响由负变正，除此之外仅有凌晨交易占比和经营稳定得分的系数为正。即：在新加入的指标相等时，交易总量越大、经营越稳定的商户会不活跃。
- 注：测试集上AUC为 90%。



谢谢  
Thanks