基于大数据的金融企业风险控制:以百融金服为例

毛基业

中国人民大学 商学院

2016-2-26

大数据之于金融企业的价值

通过解决金融企业的信息不对称，大数据可以帮助金融企业解决由于信息不对称而带来的营销、定价、风险、欺诈以及催收等问题。

大数据时代，金融机构之间的竞争将在网络信息平台上全面展开，说到底就是“数据为王”。谁掌握了数据，谁就拥有风险定价能力，谁就可以获得高额的风险收益，最终赢得竞争优势。

案例背景（www.100credit.com）

百融（北京）金融信息服务股份有限公司(简称“百融金服”)，是一家专业提供**大数据金融**信息服务的公司。

- 百融金服依托大数据技术及来自互联网、金融机构、线下零售、社交、媒体、航空、教育、运营商、品牌商等多维数据源，创新性地为信贷，保险，投资理财等行业企业提供获客引流、精准营销、客群分析、风控管理、反欺诈、贷前信审、贷后管理等服务，提升金融行业整体运营管理水平。
- 百融金服坚持开放、互补的数据联盟战略，致力于运用新技术、新手段，为金融机构搭建营销与风控体系，立志成为国内金融领域最大的第三方风控及营销服务提供商。

线上线下融合的大数据金融建模及实践效果

01 传统的风险与营销建模思路

02 线上线下融合的大数据风险与营销建模思路

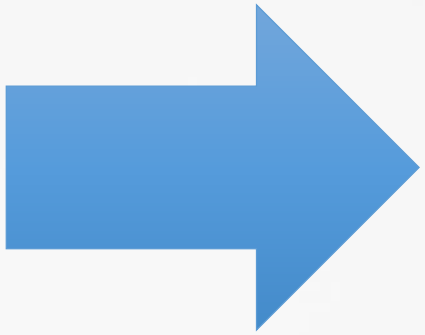
03 实践效果

01.传统的风险与营销建模思路

传统的信用评分卡变量

采用的变量：

- 信用记录时间
- 信用额度
- 借款逾期记录
- 房屋按揭还款记录
- 用款占信用额度的比例
- 坏账记录



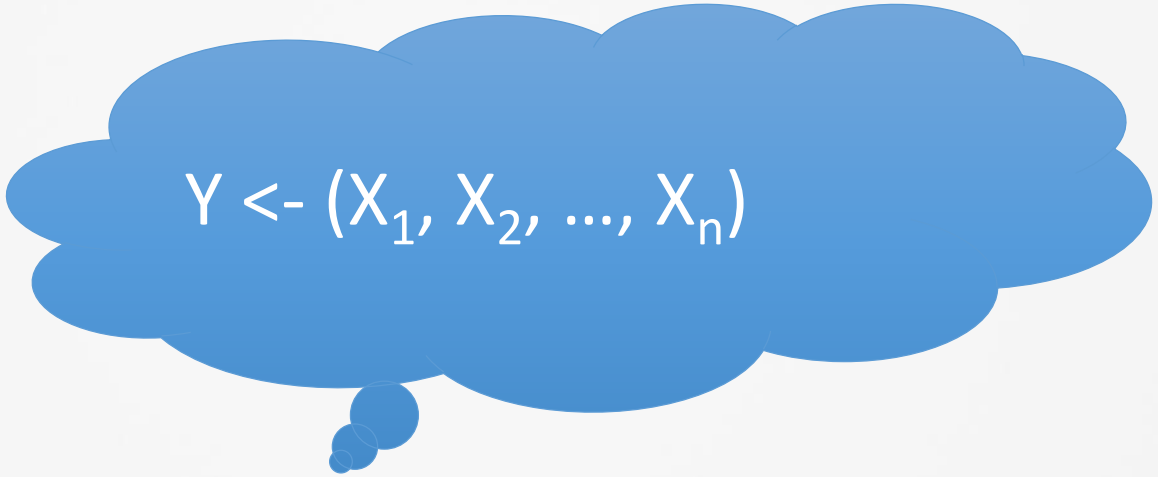
之前没有与金融机构发生借贷关系的用户无法被金融机构有效地评估信用

最重要的变量：各种还款逾期、坏账数据

传统的信用建模方法

金融机构的方法：使用金融数据进行金融建模，大概10-20个强变量

Y（还款违约概率）$= F(X_1, X_2, \ldots, X_n)$



金融类数据

中国征信体系的现状

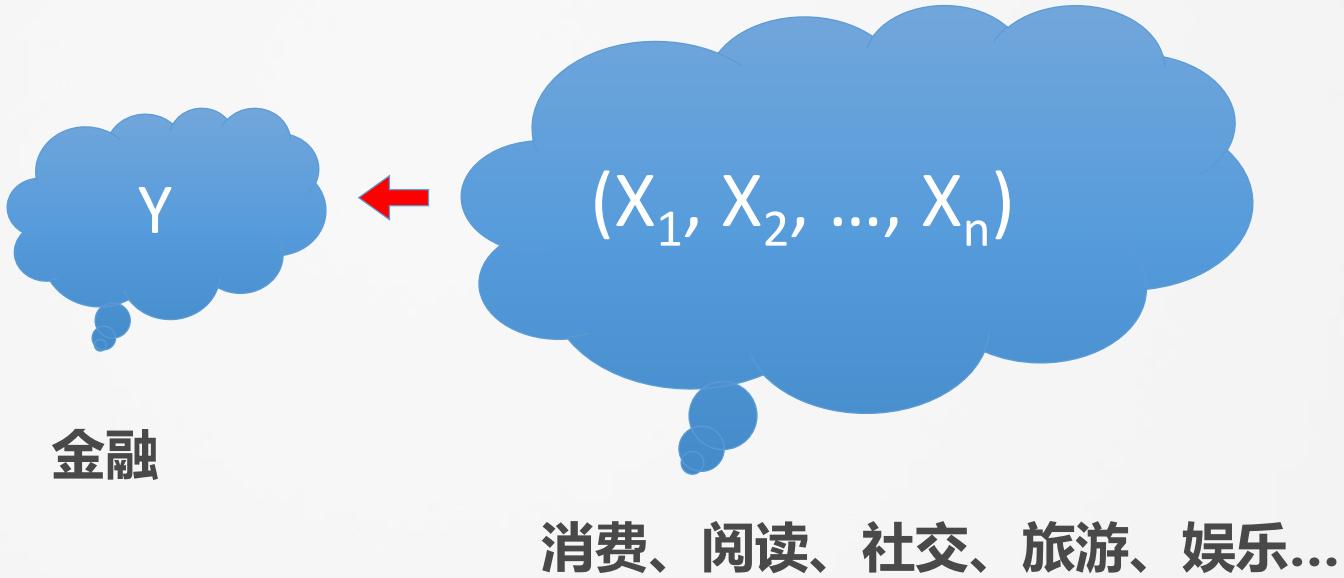
- 人民银行征信中心有3亿人的信用记录
- 占中国总人口的25%，仍有75%的人没有有效的信用记录
- 这将导致大多数人的融资需求很难得到满足

02.线上线下融合的大数据风险与营销建模思路

有否其他的信用建模方法？

百融正在尝试的方法：使用非金融数据进行金融建模，大概500,000个弱变量

Y（还款违约概率）$= F(X_1, X_2, \ldots, X_n)$



国外参考对象——ZestFinance

前Google CIO创立的大数据信贷模型公司

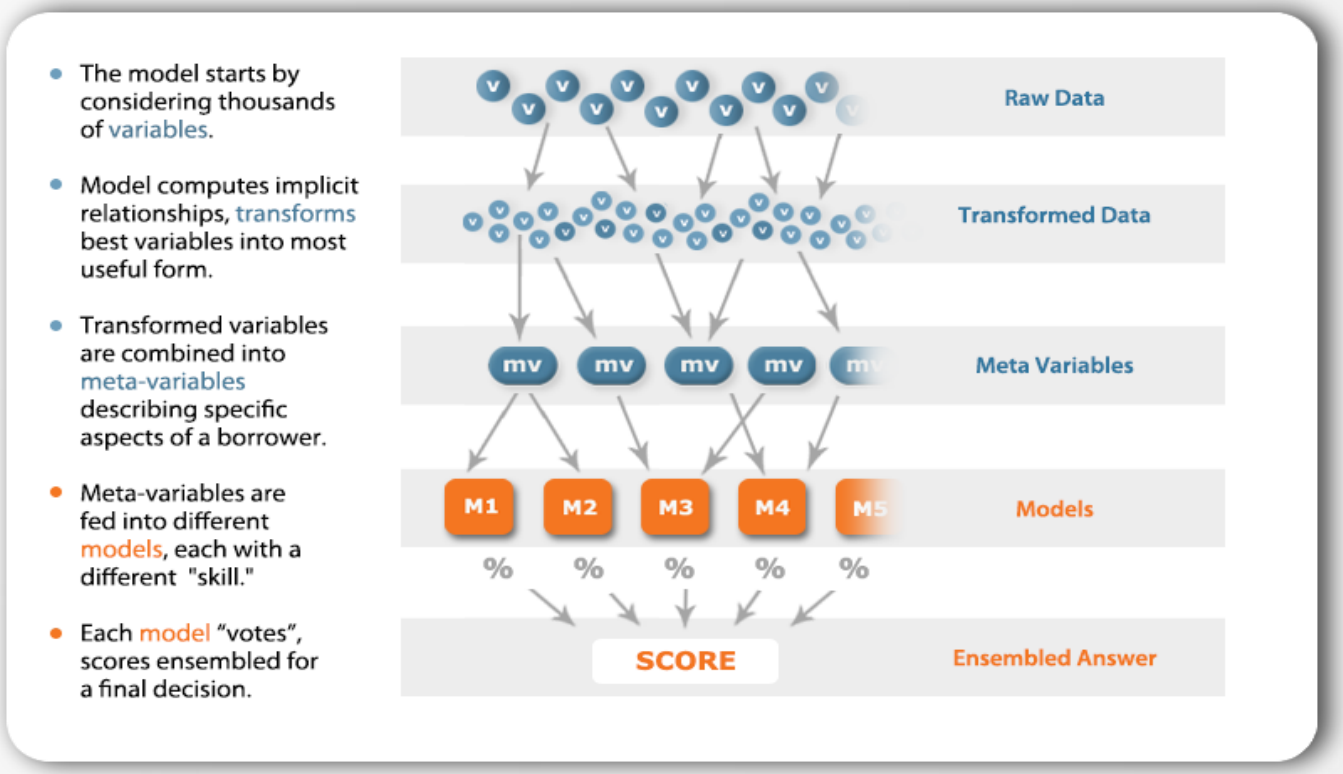
获得1亿美金投资

主要目标是帮助被FICO模型拒绝的人群重建信用以获得贷款

使用了包括以下数据来建模，

包括70000多个变量

- 用户自己填写的人口信息
- 在放贷公司网站上面的行为
- 互联网数据，尤其是SNS数据
- 与第三方公司合作得来的数据



线上线下大数据模型在风险防范上初显效果

某大型股份制商业银行A—信用卡风险评估

- 经过2轮共130万真实用户的测试，基于百融用户评估报告，可以将该行信用卡不良率降低至之前的1/2(线上+线下)

某大型股份制商业银行B—信用卡风险测试

- 经过2轮共50万真实用户的测试，基于百融用户评估报告，可以将该行信用卡不良率降低至之前的1/2(线下)和1/3(线上)

某大型股份制商业银行C—信用卡风险测试

- 经过1轮共30万真实用户的测试，基于百融用户评估报告，可以将该行信用卡不良率降低至原来的1/1.6 (线上+线下)

某业内领先的P2P公司—信用风险测试

- 线上数据整体匹配率66.77%，线下数据整体匹配率43.50%。可以将不良率降低到以前的1/2 (线上+线下)

某移动端小贷公司—欺诈及信用风险评估

- 基于百融用户评估报告，已经将该公司不良率降低至原来的1/3 (线上)

大数据风控核心思路

欺诈风险防范：真实身份识别是核心

- 欺诈客户一开始就是恶意的
- 欺诈客户很多时候不会采用真实身份来申请

信用风险防范：行为数据挖掘是核心

- 申请阶段不是恶意
- 还款能力（经济实力）与还款意愿（道德风险）较难判断

大数据云决策风险控制平台

业务支持

- 营销策略
- 反欺诈
- 准入授信
- 辅助审批
- 催收预警

云决策平台

- **规则引擎**（灵活自动化配置、预制超过300条风控规则）
- **黑名单库**（超过2000W法院执行、黑名单,包括银行、P2P、小贷等）
- **指标库**（超过2000个有效变量）
- **数据建模**（国内最先进的大数据建模团队）
- **评分报告**（总分300-1000分的信用分析报告）

云端大数据计算平台

数据来源

身份类数据
- 姓名、 证件号
- 公安信息
- 户籍信息
- 家庭成员信息

社交类数据
- 社交APP数据
- 媒体阅览数据
- 通讯运营商数据
- 网络评论数据

行为类数据
- 交易消费行为
- 网络浏览行为
- 通讯类信息
- LBS位置信息
- 航空酒店类差旅

信用类数据
- 人行信用报告
- 学历学位信息
- 公共缴费信息
- 法院纠纷判决
- 金融机构黑名单

履约能力类数据
- 银行存款
- 可信收入
- 股票投资
- 不动产类

共债类数据
- 信用卡逾期
- 本机构申请次数
- 申请过的机构数
- P2P申请次数
- 小贷申请次数

合作机构

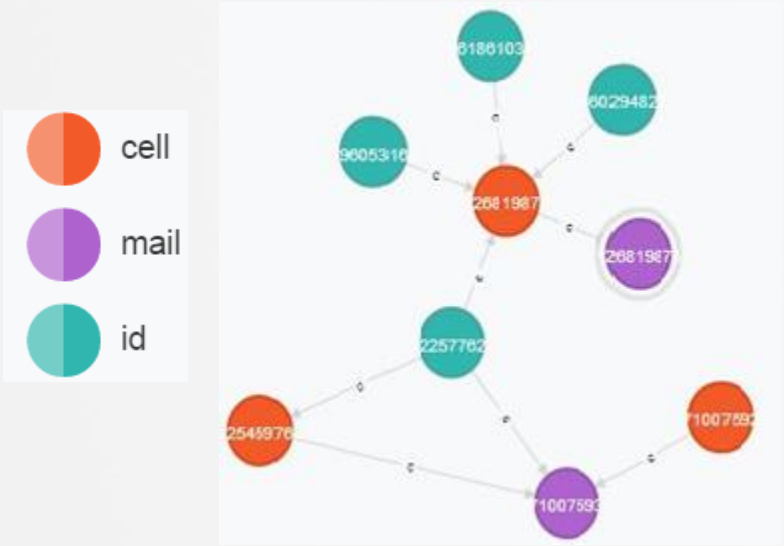
- 超过2000家电商
- 超过500家媒体
- 超过30家银行
- 超过200家P2P
- 其他合作机构包括工商、电信、公安、公共服务机构等

欺诈风险防范

欺诈风险评估规则举例

百融过去6年织下了一张关于用户各种ID对应关系的“天罗”（线上）“地网”（线下），要骗过这张网，不容易

- 实名ID：姓名、身份证号
- 准实名ID：手机号、电子邮箱、地址、银行卡号、车牌号……
- 匿名ID：QQ号、微博号、各种其他网名、浏览器Cookie、设备指纹（PC或手机硬件设备号）……

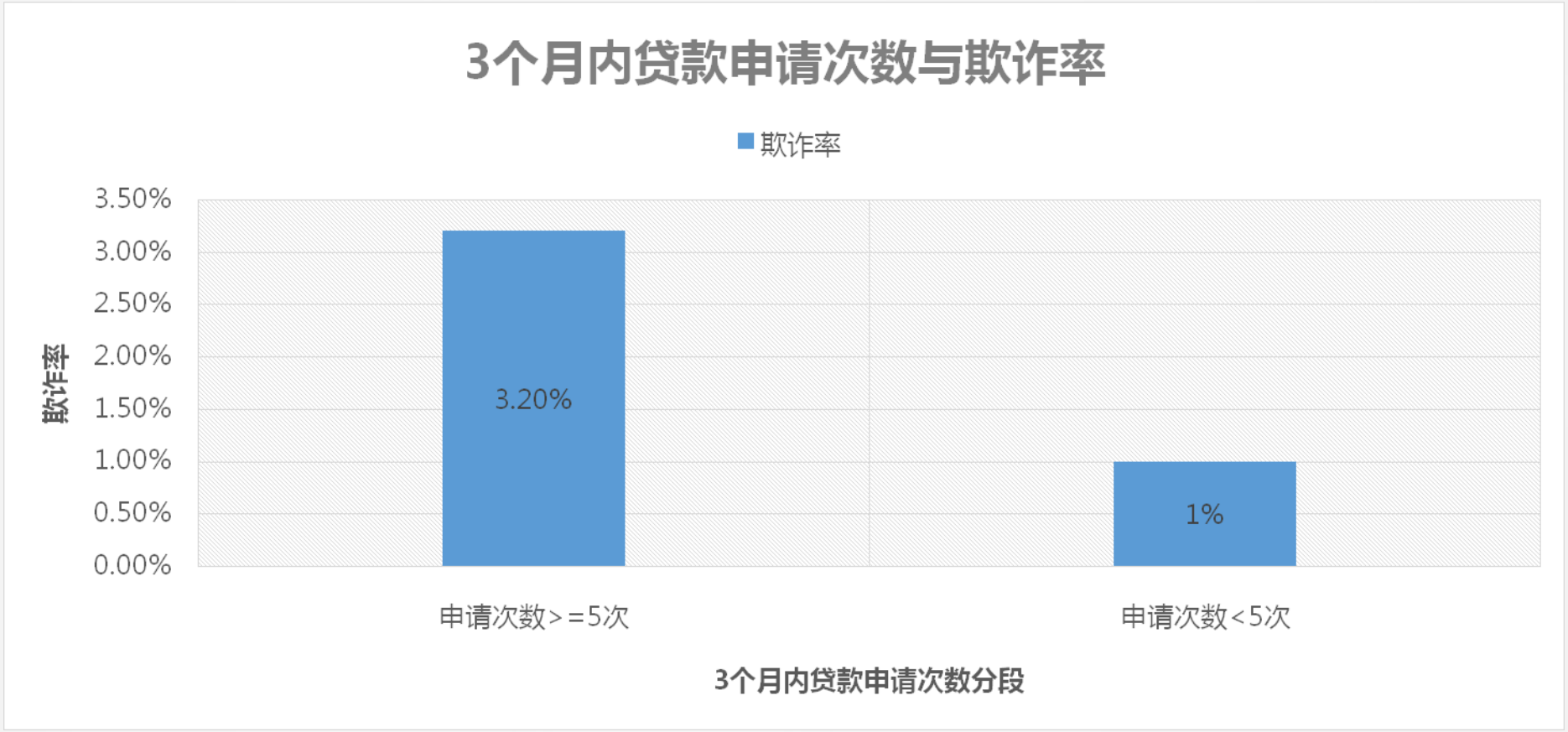


欺诈风险防范规则举例

- 同一手机在一段时间内多次申请贷款，存在欺诈嫌疑
- 同一手机在一段时间内在多家机构申请贷款，存在欺诈嫌疑
- 申请人在一段时间内更换过多个手机号或地址，存在欺诈嫌疑
- 申请人填写地址与实际居住地址差距非常远，存在欺诈嫌疑

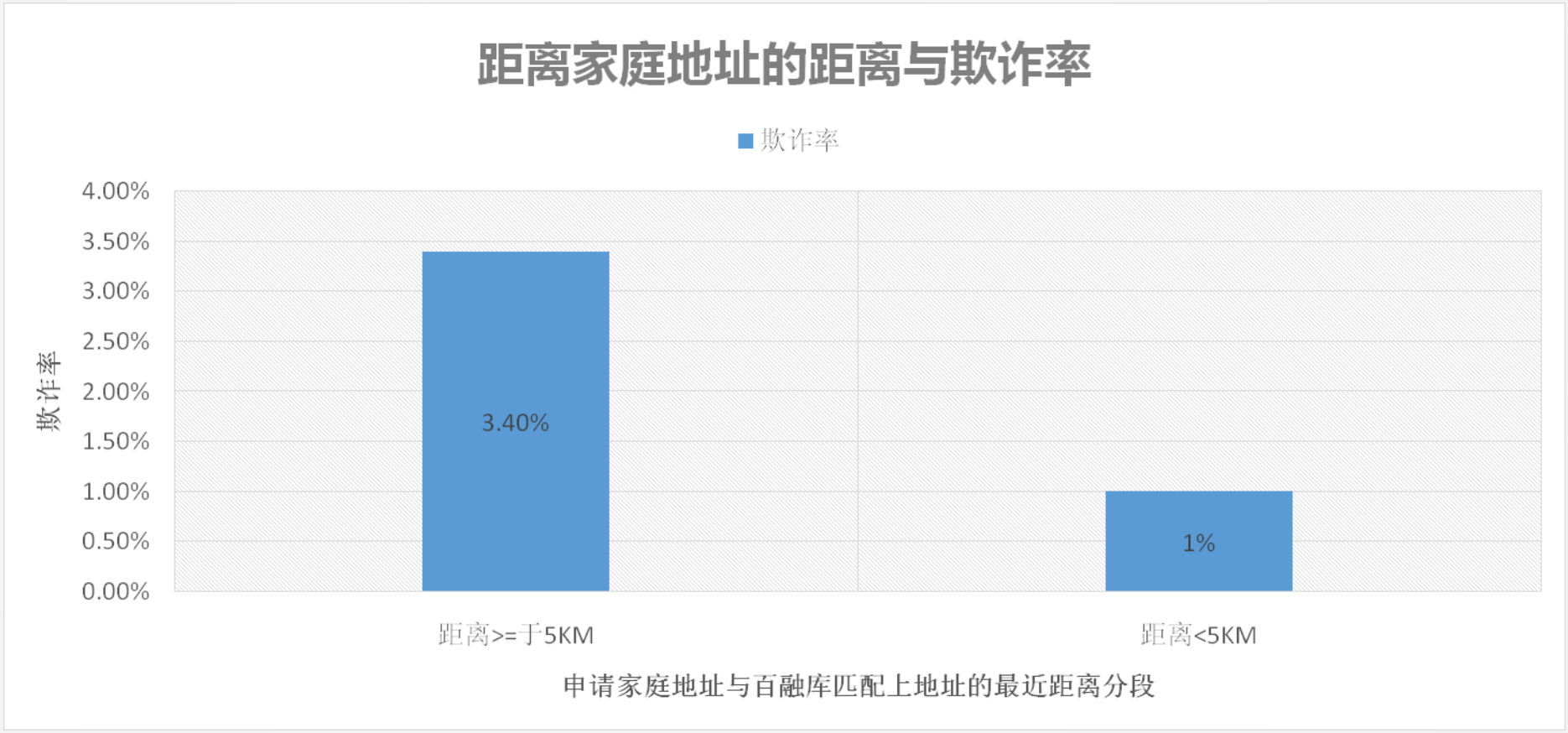
贷款申请次数与欺诈风险的关系

- 3个月之内申请过至少5次（不管是在一家机构还是多家机构）的申请者，欺诈率是其它群体的3.2倍。



地址距离与欺诈风险的关系

- 申请家庭地址与百融库匹配上地址的最近距离大于5公里，欺诈客户数是距离小于5公里欺诈客户数的3.4倍。申请家庭地址与百融库匹配上地址的最近距离越大欺诈风险相对越高。



信用风险防范

信用风险评估规则举例

要预测中国绝大部分人的还款能力与还款意愿，更多地需要依赖于分析金融行业之外的海量用户行为数据（弱变量），从中挖掘出具有可以多次复用的规律。

信用风险评估

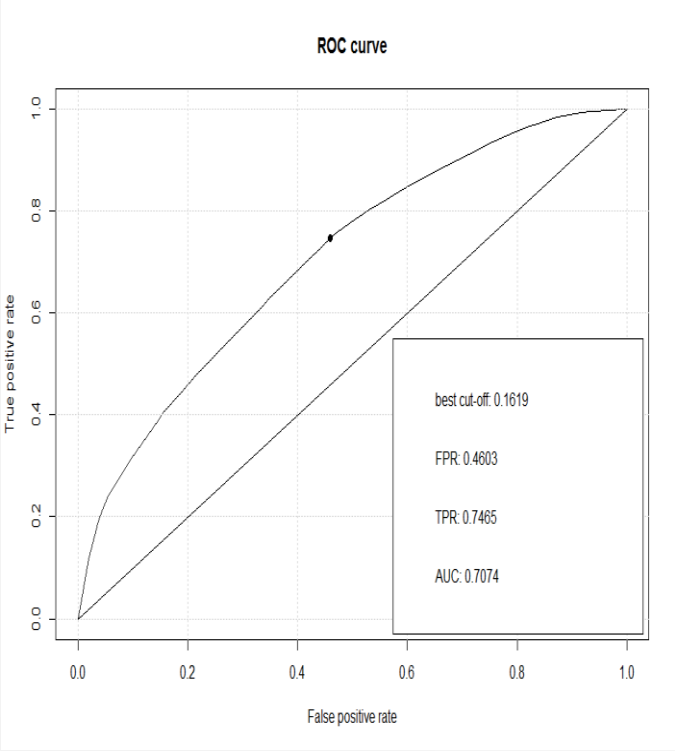
- **还款能力有限。**消费水平与收入水平不匹配，且借款未被用来提升自己的收入水平：
  - 2-4线城市、在游戏、动漫等类目消费水平过高的用户，不良率偏高
- **还款意愿较强。**受教育程度较高、道德水平较高：
  - 在财经、管理、科技等类型媒体上活跃度较高的用户，不良率偏低
- **还款能力与还款意愿都较强**：
  - 坐过商务舱以上、或者一年乘坐飞机不少于4次的用户，不良率较低
- ……

信用风险评估模型(某银行信真实模型)

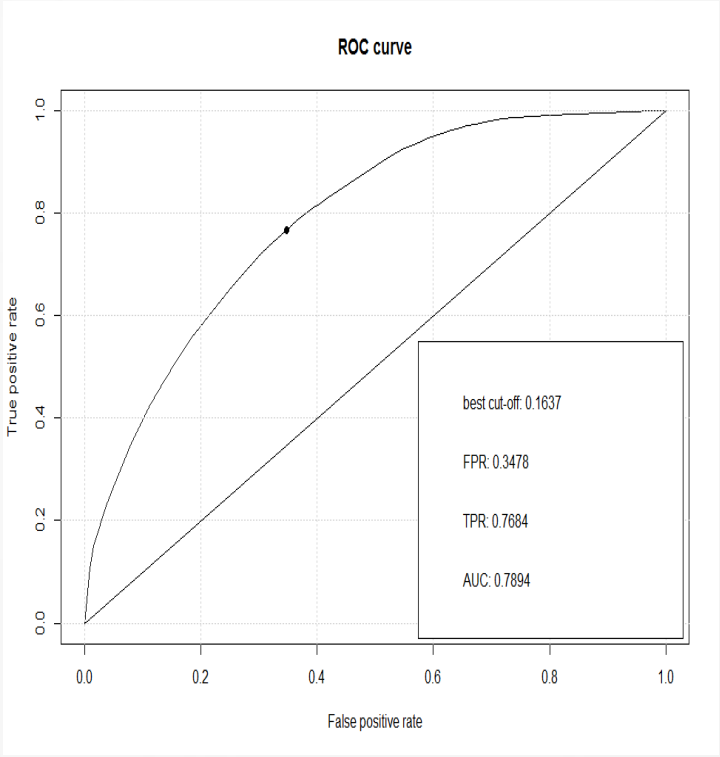
假设模型预测一批客户会产生逾期，其中a%预测正确（的确产生了逾期），b%预测错误（事实上没有逾期），则：KS = a% - b%

*注：KS值被用来评判模型区分好坏客户的能力，是银行界统一使用的标准。KS值越大模型越好

银行申请表数据
KS:0.28



银行申请表 + 百融用户评估报告
KS:0.49



KS值

申请表 + 百融报告
0.49

银行现状
申请表 + 人行报告
0.38左右

模型可用
>0.32

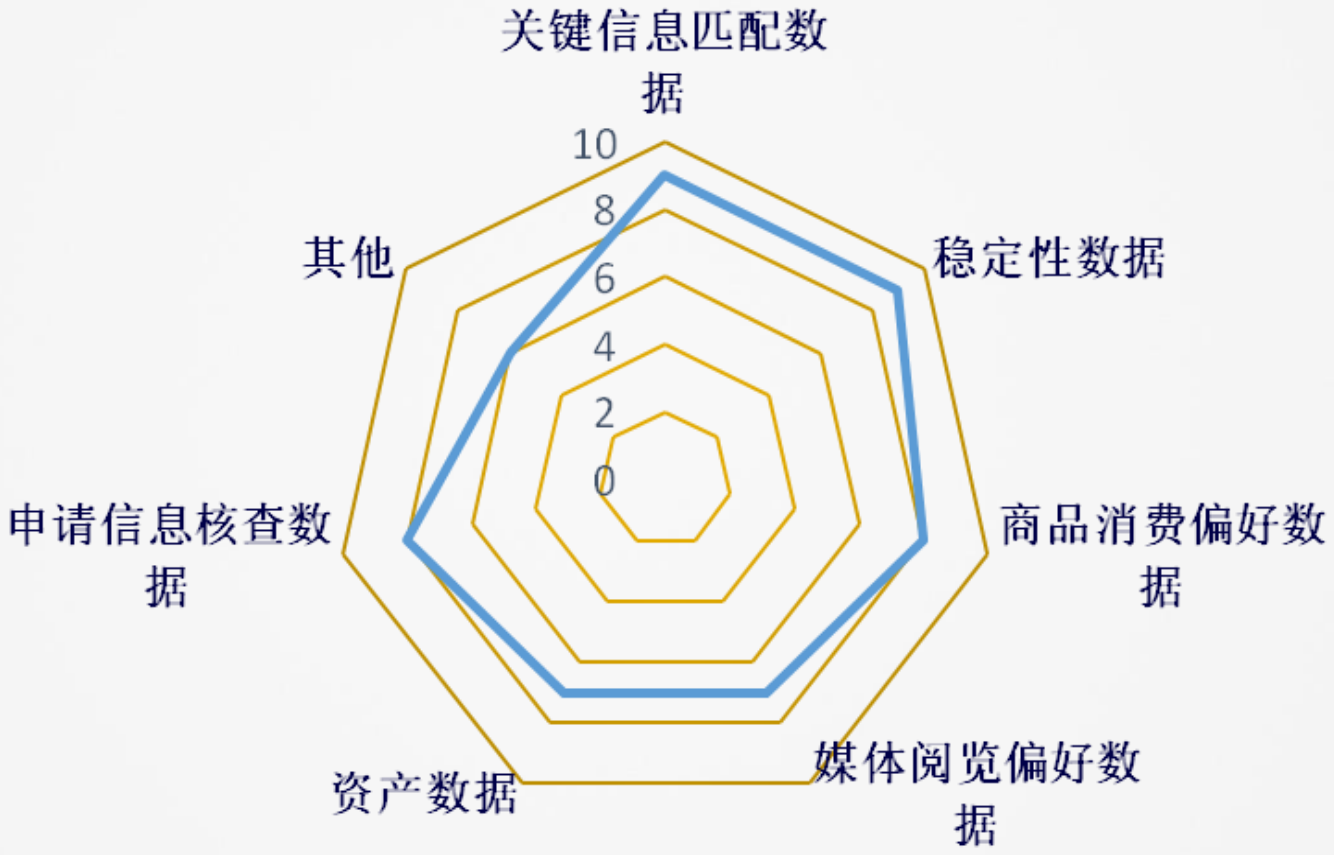
只使用申请表数据
0.28

百融评分介绍

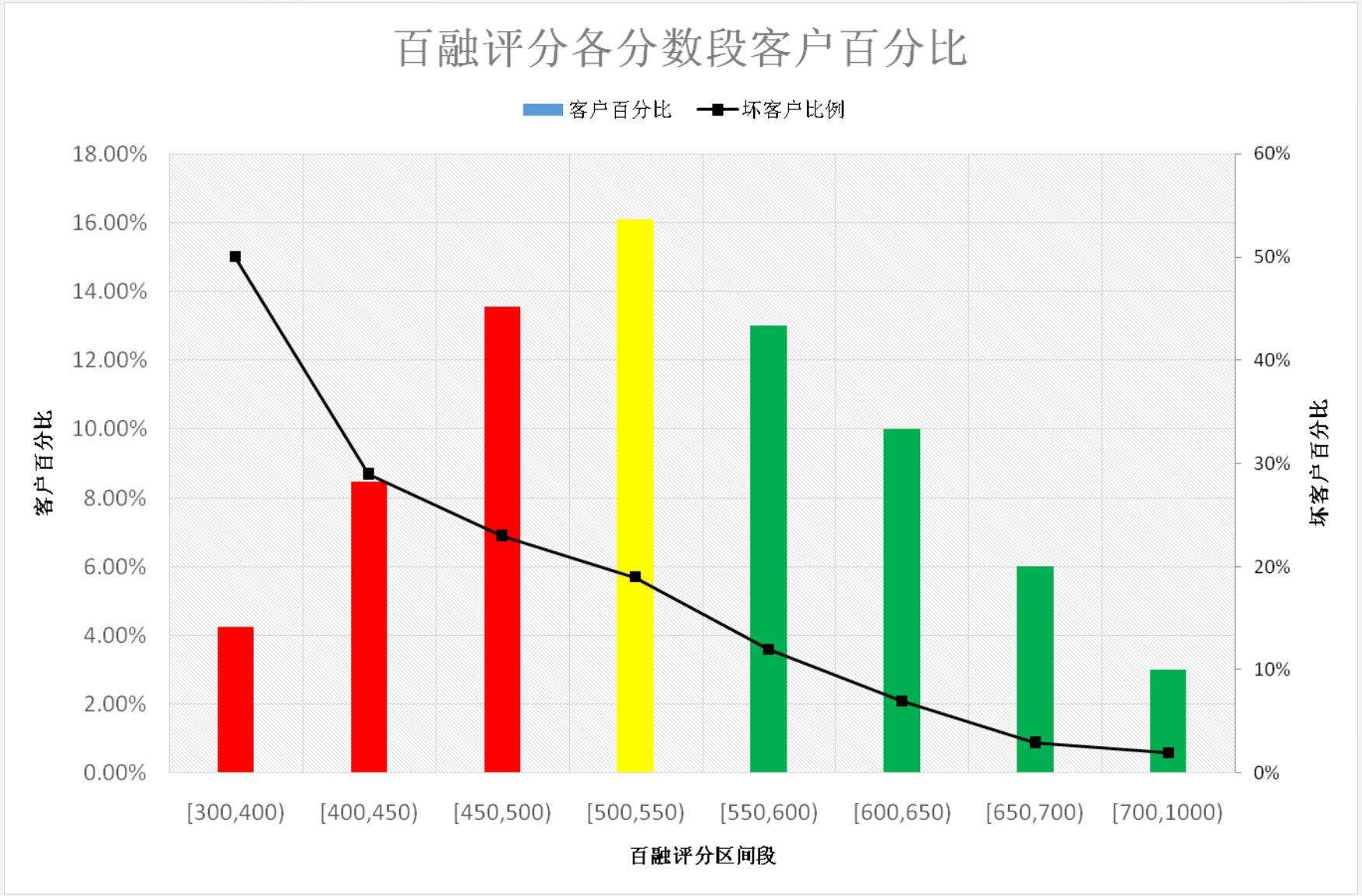
- 评分模型是基于真实的贷款违约数据建出来的
- 评分即“百融信用申请分”，将来会拓展到百融评分体系，涉及信用申请分、信用行为分、欺诈评分、催收评分等；
- 信用申请分主要基于个人最客观的行为偏好数据，利用机器学习和大数据技术，从几千个原始的弱变量中提取出能够有效识别好坏客户的强变量，再运用国际上流行的个人信用评分模式，以使模型具备有效性、稳定性和高预测能力；
- 信用申请分，在保证数据真实、客观、全面的前提下，综合评估了百融库的关键信息匹配数据、稳定性数据、商品消费偏好数据、媒体阅览偏好数据、资产数据、申请信息核查数据等，以更加准确的评价个人的信用风险。

百融评分介绍

百融评分参考因素



百融评分



- 评分分值在300~1000分之间，分数越高代表信用风险越低，违约的可能性越小。
- 不同的评分区间与信用风险等级对应如下：
  - [300,500)　高风险
  - [500,550)　中风险
  - [550,1000]　低风险

申请评分卡 - 评分卡示例

入选变量	变量取值	信用分值
发生交易次数最多的项目类别	**娱乐、游戏、动漫** 3C 数码 **餐饮食品、出差旅行** 穿衣打扮、居家生活	36 46 64 51
发生访问次数最多的项目类别	**娱乐、游戏、动漫** 时政新闻、文学/艺术、财经、生活/地方类社区、历史/社会/人文 **科学/教育、财经/管理**、汽车、旅游/交通、健康/医药、母婴育儿 知识/问答、游戏/动漫	33 48 56 40
匹配类型	只有ID或者邮箱匹配 ID和邮箱匹配 or 只有手机匹配 ID和手机匹配 手机和邮箱匹配 or ID、手机和邮箱都匹配	35 39 45 49
用户在娱乐类目的消费级别	<0.48 or 未知 >=0.48 and <0.68 >=0.68 and <0.96 >=0.96	69 56 52 50
用户3C数据类目的消费金额	<550 or 未知 >=550 and <4400 >=4400	39 51 54

数据安全资质保障

- 已经获得公安部**信息系统安全等级保护二级资质**，符合征信法规对征信企业的安全性资质要求
  - 365天每天都有人值班，每天都向公安系统发送安全报告
- 正在申请
  - 公安部**信息系统安全等级保护三级资质**：与银行同等级的资质
  - ISO 27000（国际标准化组织信息安全管理体系）
- 参照国际大数据公司的最佳实践，对关键数据进行加密、脱敏、分块管理和传输加密，确保了用户身份信息的私密性和安全性
- 自建了包含1000多台服务器的云平台，公司运营5年多以来，拥有1000多家客户。每天处理8000多万用户的行为数据挖掘任务。从未发生过一例安全事故和安全投诉

谢谢！