

大数据风控用了什么模型？有效性如何？

风控具体指的是什么？

风控中得数据、特征、算法分别有哪些？

具体的风控应用案例有哪些？

风控的评估效果如何？

风控过程中得注意点？

本题已加入知乎圆桌 » [金融科技浪潮](#) ，更多「金融科技」话题讨论欢迎关注

关注问题

写回答

添加评论

分享

邀请回答

收起

37 个回答

默认排序



项亮

104 人赞同了该回答

目前贷款的风控因为每一个样本的收集都需要放款来收集，想想每人放一万，一个亿也就只能放1万人，所以样本量不会太大。所以所谓大数据风控主要是大在特征的数据上。很多时候是用了很多传统上不怎么敢用的特征。比如传统风控比较害怕missing value 比较害怕不稳定的特征 这些都是大数据风控需要解决的。

说到模型，既然是特征多，样本少，那就需要一个非常抗过拟合的模型。另外如果是单独针对反欺诈而不是信用，因为问题比较非线性，所以需要有一个有非线性能力的模型。满足这两者要求的都可以。

当然上面说到的只是针对预测贷款用户好坏的二分类问题，至于很多风控领域的其他问题，就有不同的解决方案了。

说到有效性。据我所知目前市场上有一些非常小额短期的产品已经可以完全按照一个模型放款并盈利了。完全不需要人参与。这类产品通过小额解决了样本少的问题。通过短期解决了收集label慢的问题。所以还不太容易推广到大额长期产品上去。

编辑于 2015-11-15

▲ 104



● 11 条评论

➦ 分享

★ 收藏

♥ 感谢



黄席盛

一生只做正确的事

131 人赞同了该回答

我理解，目前大数据风控主要分为三类：

1. 反欺诈模型
2. 二元好坏模型
3. 资产包风控模型

一、反欺诈模型

大数据风控只能用于小微资产（现金贷、消费贷、小微企业贷），而不可能用于基建、政信。对于小微资产，还款能力不是核心问题，主要风险是还款意愿。因此目前市面上大数据风控90%的价值在于反欺诈。

反欺诈的大数据风控主要基于两套工具：交叉验证、聚类分析。

交叉验证主要是由人工判断规则，系统校验是否符合实际情况。如通讯录和通话记录校验、电商记录校验、设备指纹校验、多信息源地理位置校验。以现金贷产品为例，大多数现金贷产品的基础风控逻辑就是两个摄像头，后摄像头识别身份证，前摄像头做人脸的活体识别，人脸对上身份证，就做好了反欺诈，之后就扔到二元好坏模型做评估。

聚类分析和交叉验证的区别是，交叉验证很多时候根据一些人工的规则，但是聚类分析主要是根据结果反向推导。比如通过历史资产的履约情况，发现在19-25岁区间的人群风险较低、发现输入地址时间比较长的人群风险较高、发现填写收入在30000以上的风险比3000以下还高。有的规则最后可以通过逻辑解释，有的规则最后根本也无法理解为什么。但是如果一个新的进件，和之前的「坏客户」比较相似，那么他大概率是坏客户。

以同盾为例，主要向资产、资金、支付、场景四方输出反欺诈SaaS，提供：

1. 交叉验证工具

2. 聚类分析报告
3. 黑（灰）名单数据库

二、二元好坏模型

二元好坏模型的核心价值是量化定价，包括授信额度、贷款期限、利率等。主要工具就是评分卡，先给用户信用评分定级，然后不同级别不同利率。宜人贷分为ABCD类客户，利率分别为17%、27%、34%、40%；Lending Club分为从A1-G5共35个级别，利率水平从6%到26%不等。（16年初数据）

至于贷款额度，一般随行就市。

1. 个人现金贷：小额现金贷以随行就市为基础，通过拍脑袋决定，在1000-5000不等。
2. 个人消费贷：由于中国居民杠杆率较低，基本上3C、医美、教育的资金需求都低于客户授信额，因此直接使用交易金额就行。对于车贷行业，一般也是简单分档，30万以上车审核较严，10万以下车分36期，客户还款压力也不大。
3. 小微企业贷：目前大数据应用不多，主要因为小微企业造假动力强，基础数据都难以确保真实性。目前小微企业还是以抵押贷款、法人贷款、供应链融资为主，信用贷主要还是依靠IPC方式通过线下业务员重制报表实现。电商类企业的风控模型基本上是根据流水的比例来。

三、资产包风控模型

上述都是基于单笔资产的方法论，但是从资产包层面的风控有不同的考虑。

假设还款是1，逾期是0，不同的客户有不同的表现：

1. A：1111111111
2. B：0000000000
3. C：0011011101

A是好人，B是坏人，这两个问题没有异议。很多时候，基于前两种模型我们会认为C是坏人，但是从资产包层面，他提供了不菲的罚息收益。

此外，资产包的风控还要考虑不同资产的相关性，考虑优先劣后配比后的预期风险改变，考虑流动性的风险。

四、目前的市场格局和问题

第一个问题，长尾征信公司的价值。

放贷市场是碎片化的，但是征信服务提供商有规模效应，应当是集中的。也就是百融同盾两家争天下，芝麻信用、腾讯信用作为两个数据库对外输出和输入数据。

我搞不懂，在one or zero的市场环境下，为什么现在冒出那么多小的征信公司，还拿到融资，商业价值在哪里？尤其是像某些单一数据源的征信公司，我感觉被收购的价值都没有，大公司不如坐等你死然后收编团队？这个问题我没有答案，向各位专家请教。

第二个问题，过拟合问题。

信贷是周期性的，大周期小周期一堆。科技也是有周期性的，学生贷火起来，所有公司干学生贷，2年吃完整个市场，其他任何资产都面临创业公司蜂拥而上的局面。

数据量有限的情况下，模型可能过度地学习训练数据中的细节和噪音，以至于模型在新的数据上表现很差，这意味着训练数据中的噪音或者随机波动也被当做概念被模型学习了。而这件事，在市场环境发生变化之前可能没有任何人知道。

第三个问题，系统性风险。

目前大数据风控应用最广的是小额现金贷，因为他的数据反馈快（30天一反馈），因此比较容易做机器学习。市场上所有现金贷看下来，坏账率约为4-8%，都是一开始8%或者更高，通过机器学习降低到4%左右。但这个数据其实意义不大，依然无法反驳复贷的担忧：现金贷的借款人重复借款，本质上每个借款人都成为一个小的庞氏骗局池。就像当初和泛亚一起玩的经纪公司都盈利，但是最后还是免不了崩盘，过度相信科技和数据也许是金融领域更大的风险。

利益相关：没服务过征信相关项目，完全技术白痴，仅代表个人观点。

编辑于 2017-04-22

▲ 131



● 16 条评论

➤ 分享

★ 收藏

♥ 感谢

收起 ^



京东白条 ✓

数据 · 风控 · 产品 · 场景 · 用户

收录于 编辑推荐 · 129 人赞同了该回答

由于题主提出的问题围绕着风控模型，而讨论模型必定和实际的应用场景和数据源相关，因此就前四个问题一并回答。

首先金融科技公司大致分为三类，基于线上垂直领域（教育、医疗、电商）、基于特定客群（学生、蓝领、白领）、基于线下场景（车贷、租房）。不同公司在数据维度、授信客群、产品上都有较大区别。基本而言，风险主要集中于**信用风险及欺诈风险**。

在此简单介绍下消费信贷产品在贷款各个环节风控主要模型对两类风险的把控。

一、模型在信用风险的用途：

1.授信准入阶段

首先是授信准入阶段，此阶段最重要的模型是**进件评分卡模型**，数据来源主要分为申请信息、历史消费信息、外部信息（例如多投借贷、公积金等）。常用模型包含LR、Xgboost、FFM等。不同模型的选取由是否需要在线更新、可解释性、线上部署环境等多种因素决定。LR的研究非常成熟，有完整的工业分布式解决方案和在线增量学习的理论基础，包括各种带正则项的变种，是非常理想的建模方法，很多时候它还会作为基准型，用于评价复杂模型的提升效果。

一般的线性模型会遇到两个问题：

一是非线性特征的学习，比如年龄。一般使用的方法是进行变量离散化，把年龄分成不同的段或者使用稀疏编码或者自编码等算法对品类或者其他信息进行重构。

二是交互影响，例如收入特征和年龄特征的交叉。高收入的中年人是干爹和干妈，高收入的年轻人是高富帅和白富美，两者的特点完全不一样。所以我们也会使用Xgboost等模型加工非线性特征，或使用FM/FFM类算法学习交叉特征，以此提升模型拟合能力。

此外在这个环节需要注意的是，由于很多公司的数据维度是有限的，分数低的用户并不一定是逾期风险较高的，而可能仅仅是留下数据较少的用户，随着业务的逐步扩张，怎么再去找更多的维度或者在原有数据维度上构建更细腻度的特征来刻画之前无法覆盖的用户群体是关键。

其次由于黑产的猖獗，时刻需要提防刷分、养号的用户，最好的解决方式是通过分析异常群体的行为，构建有区分度的特征或者引入更多数据维度使得可以更加细腻的刻画正常用户的行为，最后还需要结合产品去完善模型。

业务扩张的时候，客群的分布可能发生较大变化，引起的概念漂移也是值得关注的。

2.用户生命周期阶段

当用户准入后需要进行用户生命周期管理，常用到模型是**行为评分卡**。

和准入阶段不一样，在这个阶段，用户由于大多已经有过至少一次的还款行为，因此可以在数据维度加入借贷数据。

除此之外，需要考虑如何调整额度和息费，保证优质的用户得到更低的息费和更高的额度，而数据表现较差的用户需要用更高的息费来覆盖风险。

但不顾风险的一味最求高收益和不求收益的低风险都是没有意义的。定价模型的重点在于对用户需求和风险的合理预估，调整各个用户群体的息费和额度档次。实则可以看成对资金在不同风险回报的分配，使得在一定的风险下，总体风险收益最大化，技术上会涉及很多带约束的优化问题。

3.催收阶段

最后一个阶段，一小部分用户会逾期进入催收阶段。

这个时期重点是失联修复和催收评分卡，即刻画用户经过一定的催收动作后还款的可能性。

失联修复很好理解，就是通过各种社交数据，建立起关系网络找出与欠款人可能相关的人或者欠款人的其他联系方式。而催收评分卡需要使用到催收数据，催收数据大多是文本音频类型文件备份，因此对这种非结构类型数据的挖掘是这个阶段的核心。

催收的时机，是催收成功最重要的因素。由于催收资源有限，我们需要按照一定的分配规则来分配催收资源。在逾期的较早时期，应该将更多的资源放在较难催收的用户上，而其他的用户可能由于是忘记还款或者其他的非恶意拖欠原因没有还钱，可能给予一段时间会自我救赎；而在催收晚期，则需要放置更多催收资源在能够催回的用户上，尽最大可能降低损失。

二、模型在反欺诈风险方面的用途：

除了上述的信用风险，还有一块较大的职责就是欺诈风险。

现阶段，业界更多关注的是有组织参与的中介欺诈，常见的如批注、盗号、薅羊毛、养号、套现等诸多行为的识别。由于是团伙作案，更多是基于社交网络的社团发现算法来对中介的识别，或者是利用套现中的地址集中性相似性等特点来识别中介，或使用时间序列算法来分析用户的历史行为轨迹，手机传感器信息等生物指纹数据来核实身份。

欺诈风险的难点有别于信用风险，在较多场景下很难定义好坏用户。因此关键在于标签的获得。通常需要同案件调查人员配合，因为他们能够准确定义欺诈，同时能够还原犯罪手法，针对于模型Y变量定义，X变量设计都很有帮助。

其次，由于对抗性强，因此如何检测未发现的欺诈模式和模型的更新速度更加关键。目前这一块工作业界发展都比较滞后。

最后，授信客群的变化或者欺诈团伙作案手法的变化导致原有模型可能失效，加上风险的滞后性，最新可用的训练数据可能已经离目前较远，如何从最新的数据获取模式与旧的数据模式的遗忘是难点。

三、补充

最后，补充如下几点模型评测的注意事项：

1. 由于线下训练环境和线上真实用户群体存在差异，模型的泛化能力很重要，需要确保模型学习到的是有区分度的模式而不是数据中的噪音。
2. 线下使用评测指标主要是刻画准确度与区分度的ks、auc、洛伦兹曲线和Lift曲线等和模型稳定性指标psi。
3. 客群逾期率的高低和公司产品的形态有重要关系，短期提升可以通过反欺诈技术得到改善、而中长期需要依托信用风险模型、但最终还得看产品的授信客群，面向不同客群的风控模型的指标对比试没有意义的。

感谢风险管理-决策智能部提供回答。

编辑于 2017-07-06

▲ 129



💬 7 条评论

➦ 分享

★ 收藏

♥ 感谢

收起 ^



iseeyou

47 人赞同了该回答

结合平时的工作经验回答下，大数据风控一般来讲有如下几个特征：

1.高对抗性

现在黑产非常庞大，刷单、薅羊毛、密码爆破、扫号、发帖机、灌水等等时时刻刻都在发生，无时无刻不在攻防。

2.灵活性

攻击者不断变化特征和行为，风控策略每天都需要更新，必须要保证风控策略的灵活性。

3.准确性

风控策略首先需要保证准确性，在保证准确性的同时再去提高召回率，准确性太低肯定会引起大量用户投诉。

大数据风控对模型的挑战：

1.模型的泛化能力

我们平时上线的一些模型，上线时可能效果非常好，但是上线后命中量基本是直线下降状态，一周后命中量可能降到接近零。不得不佩服黑产的强大，比较简单的模型意义不大，几天甚至几个小时就可以尝试出来并规避。我们知道复杂的特征和模型可以增强模型的泛化能力，采用复杂特征和更多维度的特征是很有效的。

2.模型的可解释性

风控模型识别出来的数据需要做相应的处理，任何机器识别处理都不可能完全避免用户的投诉和异议，对于模型一定要了解业务特征，能够转化为客服和用户可以理解的语言去解释，使得任何处理我们都有理有据。

3.模型的更新速度

高对抗性场景下，模型快速更新是关键

使用的模型：

1.聚类：比如常见的相似文本聚类，大量用户发相似帖子是常见的灌水行为，需要处理。

2.分类：比如我们根据已经识别的有风险和无风险的行为，去预测现在正在发生的行为，根据关键字动态去识别预测效果不错。

3.离群点检测：比如登录行为，当同ip登录大量登录失败，这种行为可能是暴力破解，当同ip登录基本全部成功，这种行为可能是机器登录，采用离群点检测发现这两类行为并处理。

4.深度学习：广告图像识别，黄色图像识别等

具体模型和技术：

我们主要使用了kmeans，dbscan，随机森林，c4.5决策树，logistic regression，cart，adaboost，svm，em，深度学习等模型。数据和特征比模型更重要，数据和特征决定了机器学习的上限，而模型和算法只是逼近这个上限而已。使用的框架有spark，storm，hadoop，caffe，libsvm，scikit-learn等

编辑于 2016-08-27

▲ 47



💬 7 条评论

➦ 分享

★ 收藏

❤ 感谢

收起 ^



穿靴子的猫

“曲线救国”，猫奴

收录于 编辑推荐、知乎圆桌 · 55 人赞同了该回答

更新一下有效性指标中的区分能力指标：

KS(Kolmogorov-Smirnov)：KS用于模型风险区分能力进行评估，指标衡量的是好坏样本累计分部之间的差值。好坏样本累计差异越大，KS指标越大，那么模型的风险区分能力越强。

KS的计算步骤如下：

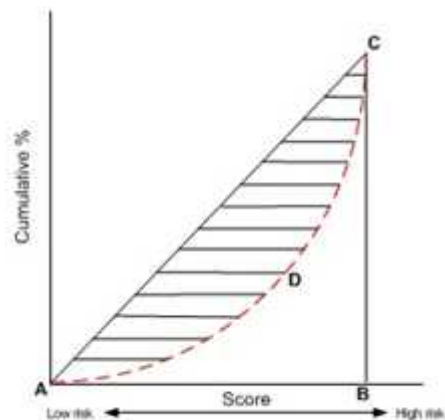
1. 计算每个评分区间的好坏账户数。
2. 计算每个评分区间的累计好账户数占总好账户数比率(good%)和累计坏账户数占总坏账户数比率(bad%)。
3. 计算每个评分区间累计坏账户占比与累计好账户占比差的绝对值（累计good%-累计bad%），然后对这些绝对值取最大值即得此评分卡的K-S值。



·**GINI系数**也是用于模型风险区分能力进行评估。GINI统计值衡量坏账户数在好账户数上的的累积分布与随机分布曲线之间的面积，好账户与坏账户分布之间的差异越大，GINI指标越高，表明模型的风险区分能力越强。

GINI系数的计算步骤如下：

1. 计算每个评分区间的好坏账户数。
2. 计算每个评分区间的累计好账户数占总好账户数比率（累计good%）和累计坏账户数占总坏账户数比率(累计bad%)。
3. 按照累计好账户占比和累计坏账户占比得出下图所示曲线ADC。
4. 计算出图中阴影部分面积，阴影面积占直角三角形ABC面积的百分比，即为GINI系数。



以下是原文

楼主范围太广。不同的行业有不同的风控目标，不同的风控过程和程度，也有不同的风控结果。其次同一行业风险也分多种风险，对不同的风险（信用风险，操作风险，市场风险）也有不同的应对办法以及模型建设。

只讲一讲中国金融行业中的银行的信用风控与大数据的渊源。

1, 风控意义与大数据建模分析优点:

中国的金融行业必定在金融全球化的洗礼下一步步找到更大市场, 相比中国制造业有成长更快的趋势。而此刻, 风控就显得尤为重要。都知道收益越大风险越大, 当然我们更想的如果是在中间找到一个平衡点让收益大的情况下拥有尽可能小的风险。而大数据建模就可以尽可能实现这点: 提高审批效率, 降低人工成本, 减少因非客观判断原因造成的失误的风险。

2, 大数据建模目标。第一点目标做信贷工厂的量化建设: 清洗银行历史数据用于数据建模形成评分卡, 再与规则结合对贷款生命周期三个阶段(申请、贷后催收)的好坏客户提供决策建议的预测框架(自动通过, 人工审核, 审慎审核, 还是建议拒绝)。

第二点目标内评合规: 背景是巴塞尔协议: 衡量银行的资本充足率和资本准备是符合巴塞尔协议的规定, 如果不符合应该采取什么样的措施。

3, 关于建模: 前: 建模的变量以及数据都是通过层层原始分析, 挖掘分析, 变量分组, 变量降维, 过度拟合VIF检测, 以及业务逻辑选择出来的。中: 而模型的建设本来有方差分析, 相关性分析, 逻辑回归, 决策树, 神经网络分析这几种。但是由于Y变量都一般为非线性所以基本都用LOGISTIC逻辑回归。后: 模型建好后还需要用PSI检验模型客群的稳定性, 用KS或者GINI函数检验模型的区分能力。(公式我就不给啦~感兴趣的孩子肯定有自己学习的方式) 如果不太理想就再改进, 这是一个做循环的闭环式过程直到选到最佳的。(PS: 建模工具: SAS, 由于可以处理相当庞大的数据且在美国极其权威的认证而著称的。别的我就不评价了嘿)

4, 好的信用风控的评估效果一主要从准确性, 稳定性, 可解释性三个方面来评估模型。其中准确性指标包括感受性曲线下面积(ROC_AUC)和区分度指标(Kolmogorov-Smirnov, KS), 稳定性指标主要参考群体稳定指数(Population Shift Index, PSI)。可解释性可通过指标重要度来进行评估, 其中指标重要度用于衡量各个解释变量对算法预测结果影响的程度。注意: 一定要将大数据建模与业务逻辑紧密联系!

分割线-----

当然, 个人觉得知道模型背后的理论也是非常有必要的。让我们顺着逻辑回归来讲。

—

首先是假设检验中假设建立。什么是假设检验呢，假设检验背后的原理是什么呢，我们模型中具体的假设是什么呢。

假设检验分为原假设 H 和备择假设 H_0 ，我们后面会推翻 H 来证明我们的 H_0 是正确的。

假设检验的原理也就是我们要推翻的这个 H 的理由是:小概率事件不可能发生。（在此我举一个经典的例子）

例 7.1.1(女士品茶试验) 一种奶茶由牛奶与茶按一定比例混合而成,可以先倒茶后倒奶(记为 TM),也可以反过来(记为 MT).某女士声称她可以鉴别是 TM 还是 MT,周围品茶的人对此产生了议论,“这怎么可能呢?”“她在胡言乱语.”“不可想象.”在场的费希尔也在思索这个问题,他提议做一项试验来检验如下假设(命题)是否可以接受.

假设 H :该女士无此种鉴别能力

他准备了 10 杯调制好的奶茶, TM 与 MT 都有. 服务员一杯一杯地奉上, 让该女士品尝, 说出是 TM 还是 MT, 结果那位女士竟然正确地分辨出 10 杯奶茶中的每一杯. 这时该如何对此作出判断呢?

费希尔的想法是:假如假设 H 是正确的,即该女士无此种鉴别能力,她只能猜,每次猜对的概率为 $1/2$, 10 次都猜对的概率为 $2^{-10} < 0.001$, 这是一个很小的概率,在一次试验中几乎不会发生的事件,如今该事件竟然发生了,这只能说明原假设 H 不当,应予以拒绝,而认为该女士确有辨别奶茶中 TM 与 MT 的能力. 这就是费希尔用试验结果对假设 H 的对错进行判断的思维方式可归纳如下.

假如试验结果与假设 H 发生矛盾就拒绝原假设 H , 否则就接受原假设.

在模型中我们的假设便是我们逻辑回归的因变量和自变量之间没有线性关系。

$$p = \frac{1}{1 + e^{-g(x)}}$$

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = g(X) = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k$$

也就是这里面的beta们都是0。

二，never say yes.在原假设正确的前提下，确定检验统计数并计算出统计数的估计值（即构造统计量并计算统计量的估计值）

一般我们会把统计量构造成符合正态分布、卡方分布、F分布的情况，由构造的统计量不同可分为u检验、卡方检验、F检验等。

这里我们以卡方分布统计量为例子：

Pearson提出如下统计量：

$$\chi^2 = \sum \frac{(\text{实际频数} - \text{理论频数})^2}{\text{理论频数}} = \sum \frac{(A - T)^2}{T}$$

χ^2 用来反映各类中实际观察到的频数与一定假设下的理论频数的偏离程度。

实际频数是通过调查或实验得到的，理论频数要按照统计假设计算出来。

在各种假设情形下，实际频数与理论频数偏离的总和即为卡方值，它近似服从卡方为V的卡方分布，因此可以用卡方分布的理论来进行假设检验。

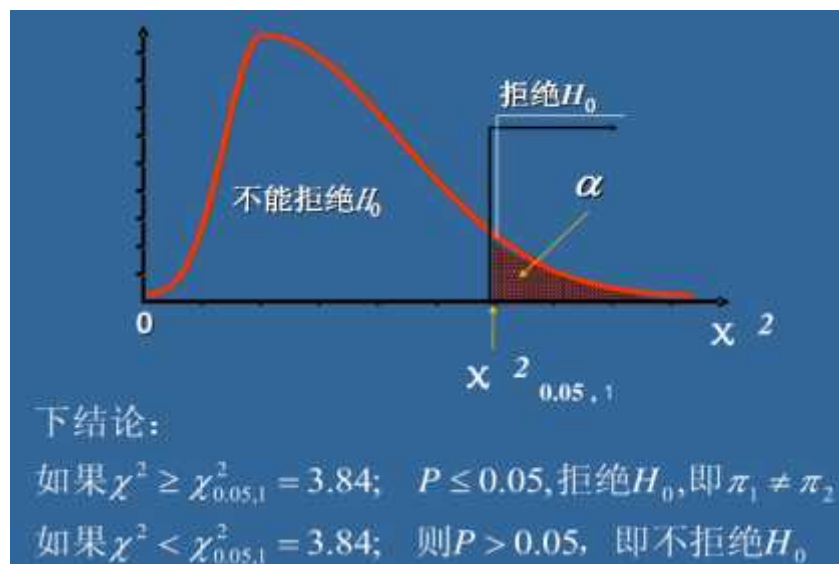
三、计算P值，或确定临界值，并比较临界值与统计数值的大小，根据“小概率事件在一次实验中几乎是不可能发生的原理”得出结论统计结果分析

显著性水平：这里的显著是一个统计学的概念，是指原假设发生是一个小概率事件，统计学上用来确定或否定原假设为小概率事件的概率标准叫做显著性水平。原假设发生的概率如果小于或等于5%，一般认为认为是小概率事件，这也是统计学上达到了“显著”，这时的显著性水平为5%。

拒绝域：当由样本计算的统计量落入该区域内则拒绝原假设，接受备择假设，拒绝域的边界称为临界值。当原假设正确时，它被拒绝的概率不得超过给定的显著性水平 α （阿尔法），阿尔法通常取值为0.05,0.01，因此落在拒绝域内是一个小概率事件。

还是以卡方检验为例

以下是卡方分布的密度函数，X轴是卡方值，Y轴是发生的P概率。



换句简单易懂的话就是，我们计算实际频数与理论频数的偏离程度即卡方值非常大的情况下概率是非常小的是不会发生的，当X2卡方值远远大于3.84，相应的我们X轴远方对应的就是越来越小的P概率。那么也就是说我们的假设是不成立的，也就是说因变量和自变量之间他们是相关的。并且在原假设情况下卡方值越大也就代表越不可能不相关，也就是越可能相关。

当然在确定检验我们单个系数的时候会用来卡方检验，整个模型的检验的时候就会用到我们F检验，T检验，他们都和我们的卡方有一定的联系。

编辑于 2017-05-22

▲ 55



💬 3 条评论

➦ 分享

★ 收藏

❤ 感谢

收起 ^



张大万

学数学的，写了本《从1开始 数据分析师成长之路》，欢迎围观。

71 人赞同了该回答

为什么觉得大家都在胡扯一气...

行业普遍使用的模型：**Logistic回归**

目前国内90%以上的建模团队都使用Logistic回归做评分卡，当然还有少数人使用决策树，神经网络和机器学习目前还没在此行业有显著成果。

Logistic制作评分卡模型的衡量标准是K-S值的大小，依据数据质量和建模能力在0-0.5之间，一般在0.3以上才可用，好的模型可以达到0.35。

芝麻分模型的K-S值在0.32左右。

以上是针对主问题给的答复，附加问题太多太散，涉及面太广，建议题主先熟悉下这个行业

以上！

发布于 2016-05-16

▲ 71



● 13 条评论

➤ 分享

★ 收藏

♥ 感谢



老夫

消费金融创业公司合伙人，曾工作于捷信

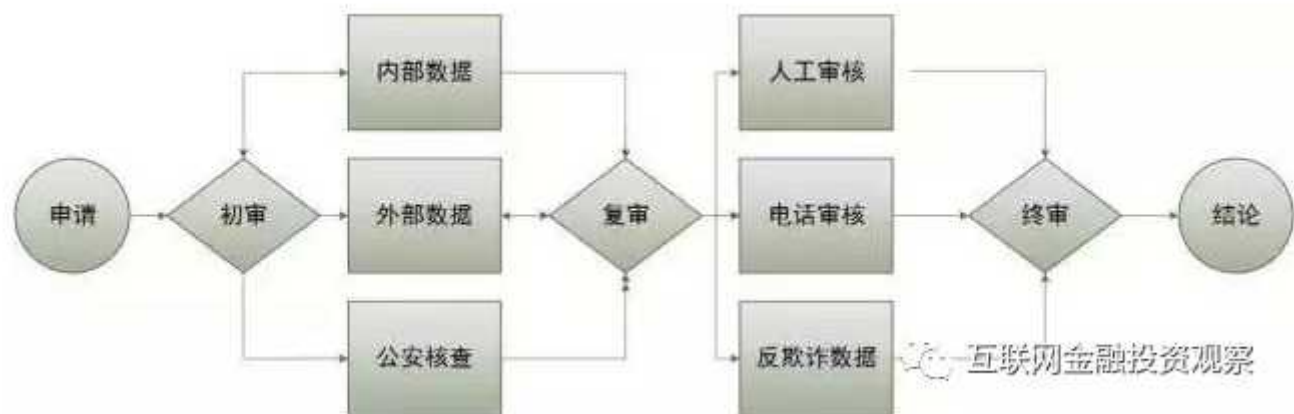
12 人赞同了该回答

相对于传统风控，大数据风控在建模原理和方法论上并无本质区别，只不过是通过互联网的红利，采集到更多维的数据变量，通过分析数据的相关性来加强或者替代传统的强因果关系。

建模原理和方法论上并无本质区别

大数据风控即大数据风险控制，**是指通过运用大数据构建模型的方法对借款人进行风险控制和风险提示。**

要理解大数据风控，首先要把传统金融风控搞清楚。这里以银行的信用卡部门为例，解析一下传统银行的信用审批流程。（附图综合了几家银行信用卡中心的审核流程）



信用卡审核简易流程图

从流程上看，银行的信用审核，是以风控评分卡模型的自动审核为主，以人工审核为辅的模式，在需要特定审核的环节由人工进行，比如验证你的工作、校验你联系人的真实性等。这也是为什么在现实生活中一部分人在信用卡申请过程中会收到人工审核电话，一部分人并不需要该验证环节即可获得信用卡。

从审核数据上看，对于银行来说，影响审批额度的主要因素包括客户基本特征（包括男女、年龄、教育程度等等）、客户的风险暴露情况（社会收入、债务情况、还债能力综合评估）、现有的社会表现（房贷还款情况、其他银行信用卡使用情况等）。

不管是中资还是外资银行，大致都遵循了这样一套风险评估和信用审核的逻辑。**对比之下，就可以看出，时下互联网金融鼓吹的大数据风控在原理和方法论上跟传统金融的风险控制并没有本质区别。**

市场空白给予机会 以数据相关性替代因果关系

大数据风控相对于传统风控来说，建模方式和原理其实是一样的，**其核心是侧重在利用更多维的数据，更多互联网的足迹，更多传统金融没有触及到的数据。**

比如电商的网页浏览、客户在app的行为轨迹、甚至GPS的位置信息等，这些信息看似和一个客户是否可能违约没有直接关系，但实则通过大量的数据累积，能够产生出非常有效的识别客户的能力。

		传统金融风控	大数据风控
不同点	数据量	传统数据 强变量	非传统数据 弱变量
	运行逻辑	强因果关系	不讲因果关系
相同点	建模规则		



大数据风控与传统银行风控的比较

数据量大是大数据风控一直宣传的活字招牌，至于多少的数据量级才能算得上大，业内一直没有统一或者较为通用的标准。根据公开资料，蚂蚁金服的风控核心CTU投入了2200多台服务器，专门用于风险的检测、分析和处置。新华网的报道显示，蚂蚁金服每天处理2亿条数据，数据维度有10万多个。京东金融2016年6月，投资了美国的大数据公司ZestFinance，之后还与其联合发起成立了合资公司ZRobot。ZRobot主要定位在为互金企业提供数据建模、信用评分、资产定价、欺诈识别等服务。京东金融依靠中国最大的电商-京东的数据量，在国内已算大数据拥有者。

聚秀资本合伙人江南愤青表示，按照惠普副总裁提及的大数据概念，全球有能力进行所谓的大数据应用的公司不超过50家。大量的公司只是在做数据的优化，根本不能称之为大数据风控。

在数据维度这个层级，传统金融风控和大数据风控还有一个显著的区别在于传统金融数据和非传统金融数据的应用。传统的金融数据包括我们上文中提及的个人社会特征、收入、借贷情况等等。而互金公司的大数据风控，采纳了大量的非传统金融数据。比如阿里巴巴的网购记录，京东的消费记录等等。

在运行逻辑上，不强调强因果关系，看重统计学上的相关性是大数据风控区别于传统金融风控的典型特征。传统金融机构强调因果，讲究两个变量之间必须存在逻辑上能够讲通因果。一位不愿具名的前城商行信用卡中心负责人表示，在银行的信用评审中，他们即便发现了一些非传统变量在统计上看来跟审核结果存在某种相关性，如果不能够在逻辑上讲通，他们也断然不会采用。

“比如我们发现在某个时间点来申请的客户，从后期数据表现上来看逾期的概率就是比较高。但如果没办法从逻辑上解释通其中的道理，我们是不会贸然把它作为因变量放在审核模型当中去的。”

但与传统金融机构不同，互金机构的大数据风控吸收的正是大量的潜在相关性数据。为何说是潜在？因为通过互联网的方式抓取大量数据之后，一定会有一个数据分析和筛选的过程，在这个过程中，大量数据会被证明不相关直接被踢掉。留下的相关性数据才会被运用到风险审核当中去。

传统的线下小贷公司在放贷过程中，会有一些自己的经验判断，在面对一些特定行为特征、生活习惯的客户会首先有一个自己的直观打分判断，这些是长期经验累积的结果。现在一些互金公司可以通过技术化的手段把这些也变成输入变量纳入到风控审核当中去。

大数据风控需要纳入非传统变量，将风控审核的因果关系放宽到相关关系是有其业务原因的。伴随着互联网金融的火热，大数据风控逐渐升温。中国的互联网金融，服务的客群简言之可以分为两类：无信贷历史记录者和差信贷历史记录者。而这两部分人群，恰恰是中国传统金融机构没有服务到的两部分人群。

这两部分人群包括中国的学生、蓝领、以及一部分的白领等。这部分客群，在央行没有征信报告，几乎没有过往金融服务记录，照搬传统金融的风险审核会出现水土不服的状况。

对传统金融机构而已，在对一个客户进行信用风险评估时，工作单位是强变量。这直接关系到他的社保记录。但对一个没有固定工作的客户来讲，工作单位就变成了一个弱变量，对于最后的风控审核助力有限。


同理，学历、居住地、借贷记录这些传统的强金融风控指标可能在面对无信贷记录者和差信贷记录者时都会面临同样的问题。这迫使互金公司需要通过其他方式补充新的风控数据来源，并且验证这些数据的有效性。

场景厮杀激烈 大数据风控有效性有待验证

相对于传统金融机构，互金公司扩大了非传统数据获取的途径，对于新客户群体的风险定价，是一种风险数据的补充。但这些数据的金融属性有多强，仍然有待验证。

而数据的金融属性取决于如何去挖掘，如京东商城上购物记录其实是目标客群很好的刻画，送货的地址，GPS经常驻留的地址等，是一个人的居住地的概率很大。在这一点上，腾讯的微众银行、京东金融，蚂蚁金服等互联网巨头手中都掌握着海量的数据。

公司	风控体系	工作机制
蚂蚁金服	CTU 智能风控大脑	CTU 的核心就是判断是不是账户主人在操作，交易请求是不是可信。判断依据就是我们所熟知的支付宝、余额宝、招财宝、芝麻信用、网商银行等业务数据。
微众银行 (腾讯旗下)	多重风控模型	通过社交大数据与央行征信等传统银行信用数据结合，运用社交圈、行为特征、交易网、基本社会特征、人行征信 5 个维度对客户综合评级，运用大量的指标构建多重模型，以快速识别客户的信用风险。
百度金融	主动预警捕捉高危行为	百度金融主要是打通“人+手机+设备+IP”等关联纬度，基于全网行为进行监测，捕捉高危行为特征，在贷前准入方面就开始排查风险，对借款人的行为进行预测。在贷款后，也会对借款人贷后行为进行跟踪和监测，只要触发预警规则，也会激发提醒。
京东金融	由多种大数据机器学习模型构成的弱分类组合预测模型	京东消费金融业务风控以京东商城庞大的交易数据为基础，同时覆盖了物流、用户等京东生态体系内的所有有效数据，开发出风险控制模型体系、量化运营模型体系、用户洞察模型体系、大数据征信模型体系。
网易金融	北斗七大风控模型	网易北斗在贷前做了获客引流模型、反欺诈型模型以及风控授信模型，先构建了筛选机制。在贷中又做了信贷管理模型，确定放贷的金额以及调查还贷能力等。在贷后还有风险预警模型、云催收模型和用户增值模型。

 互联网金融投资观察

各大公司的风控体系 来源：根据网络公开资料整理

根据《证券日报》报道，微众银行旗下微粒贷的单笔均借款金额低于1万元，逾期率低于0.3%。“微众可以拿到腾讯的数据，这是其他所有公司没发比的，在小额借贷领域，他们的优势太明显了。”前述不愿具名人士透露。

巨头优势明显，但大公司不可能面面俱到，布局下各种场景。并不代表创业公司的路已被堵死。在互联网巨头尚未涉及的领域，小步快跑，比巨头更早的抢下赛道，拿到数据，并且优化自己的数据应用能力，成为创业公司杀出重围的一条路径。

有一个稳定的场景，能够在自然状态下真实地采集到客户行为所展现的数据，这是大数据风控的前提。在一些尚未被巨头嗅到的场景领域，竞争厮杀已经非常激烈。

农分期、会分期、房司令、租房宝、蜡笔分期、学好贷、爱旅行、趣分期、分期乐、买单侠、优分期……农业、租房、蓝领、学生、旅游等各个场景和不同人群下的争夺已经日趋白热化。

陆金所CEO计葵生在2016年的中国支付清算与互联网金融论坛上自曝陆金所的年华坏账率在5%——6%。并且，根据腾讯财经的报道，计葵生指出，如果风控做不好，P2P的行业坏账率将远超10%。根据新经济100人的报道，学生分期起家的分期乐坏账率低于1%。银监会数据显示，2016年第三季度，我国大型商业银行的不良贷款率为1.67%。

而在今年11月，《21世纪经济报道》披露的苏宁消费金融公司的贷款不良率高达10.37%。该文章指出，苏宁内部人士透露，10.37%的坏账绝对不是行业最高的，很多面向大学生提供分期消费的平台，不良率超过25%。

坏账率、不良率、逾期率，各种不同的指标计算口径不同，结果大相径庭。缺乏统一的行业标准，野蛮生长下也不乏充斥着故意夸大和谎言之嫌。互金行业的坏账像一个披着面纱的女郎，始终不得其真容。

不同客群的坏账表现有其梯度差异，但是良好的数据获取和数据应用能力可能会在一定程度上优化数字表现，成为企业的一道有力护城河，这也是留给创业公司的一个机会。

国内的大数据风控困境

首先是中国征信体系的不完善。要知道大数据风控的第一步就是获取数据。波士顿咨询的报告显示，央行个人征信记录覆盖率仅仅为35%。而互金企业的目标用户也多为信用卡无法触达的人群，可想而知，这批人就更没有什么信用记录可言了。而各家消费金融公司的数据相互分享可能性很小。现在大多数公司的做法是将自己的数据共享给第三方征信机构，再从征信机构那里获取数据，但这种数据的有效性存疑。获取有用数据或许成为很多公司构建自己的大数据风控模型的第一个难题。

其次是中国的团体欺诈现象。前Capital One高管，现任趣店CRO的粘旻环女士就表示，“目前国内的信用市场，反欺诈仍然是头号难题”。在中国，这种欺诈套现早已做成了一个产业链，从中介公司到商家甚至是自家公司的销售，沆瀣一气。通过各种方式召集法律意识淡薄的用户来进行借贷，再将借到的钱瓜分。而诈骗分子跑路后，还款以及逾期都压到了用户的头上。

然而前来申请借款的用户用的都是真实的信息，平台给用户的额度也在合理的范围内，这样的诈骗方式让平台处于很被动的处境。现在的处理方式只能是发现一起就抓一起，发生之后处理的速度是关键。不过粘旻环女士也表示之后会采用更主动的方式来防御。“目前，我们在搜集我们自己和

同行们遇到的相关案例，寻找这部分容易被利用的人群身上的共性。在有足够的样本以后，我们可以梳理出这些用户的画像，并建立相关的风控模型。”

第三个难题就是金融行业频发的“黑天鹅”事件。如今大数据被吹的神乎其神的一个重要原因就是认为它可以有效地推演及预测未来。但是立足于统计学基础之上的大数据可以预测出跳出规则之外的黑天鹅事件吗？恐怕很难。在国内大数据风控的发展仅仅经历了几年的时间，在这期间中国还未发生过类似2008年美国次贷危机的大规模金融危机。因此，国内大部分公司构建的大数据风控体系没有经历过极端经济环境的压力测试，届时可能完全失灵。

做风控审核，其实是审人，人性的展现，大数据模型虽然讲究的是大和相关性，但用于金融的风控，有些前提是必要的：

1 这些个大数据必须是客户自然行为的流露和展现，这样才能避免逆向选择，数据才有效。

2 采集的过程稳定，可持续，这样才长久。

3 数据够一定厚度，才能真正起到作用。

现在的阿里腾讯京东做金融大数据风控都有如上一些特点。

编辑于 2017-04-26

▲ 12



● 添加评论

➤ 分享

★ 收藏

♥ 感谢

收起 ^



神马大仙

21 人赞同了该回答

第一次用知乎回答问题，回答的不对的地方欢迎指正。

一、什么是风控，具体指什么？

很多行业会用到风控这个词汇，像券商、保险、银行甚至制造业，都会设置风控这个岗位，风控的意义是通过各种手段去管理可以预见的风险，保证公司业务的收益。

本人做的是个人贷款风险建模，对其他行业不是很了解，这里主要讲个贷。个人贷款的风控具体通过反欺诈、信贷策略、审批、贷后管理手段保证贷款本金和利息能够收回。

二、用到的大数据有哪些，获取渠道？

目前用来建模的数据包含：

- 1、申请表数据（身份信息、收入水平、工作单位、联系人等），这部分是申请贷款时客户自己填写的。
- 2、行为数据（消费能力、地理位置、购物偏好等），这部分是通过客户授权采集到的。
- 3、信贷历史（信用卡数量、还款历史、房贷信息等），这部分是央行征信查询获得的。
- 4、行内数据（存款额、卡数量、用户等级等），这部分是存量客户在某行存款、开卡等记录。

三、应用案例

- 1、欺诈风险用到模型主要是社会关系网络模型，通过每笔案件之间的关系，判断新案件是欺诈申请的可能性。
- 2、信用风险主要用到模型是逻辑回归建立评分卡（也有的用决策树），量化新申请人可能违约的概率，根据评分高低制定不同的授信规则和催收策略。
- 3、贷后管理也用到行为评分卡，例如额度调整和客户风险分池管理等。

现在很多金融机构都能够用大数据模型自动做出决策，大数据风控也用到很多场景中，比如租房分期、手机分期、二手车等。前几天本人刚刚通过某知名房产中介的app，贷款租了一套房子，全程无人工审核，两分钟搞定，非常便捷。

四、评估效果

模型效果评估指标大同小异，KS值，GINI系数，ROC等都是评价模型区分好坏客户的能力，以目前我国数据质量来看，一般来说KS达到37以上就不错了。再一个是人群稳定性PSI指数，当客户趋于不稳定时，就该重新调整风控策略，或者重建评分卡了。

五、风控工作的注意点

每个行业不一样，单以个贷来说，每家公司的优势都不尽相同，目前做消费贷款的公司既有企业，也有网商，还有银行系。这个问题真是不好回答，只能说八仙过海各显神通吧。

发布于 2016-08-28

▲ 21



💬 5 条评论

➦ 分享

★ 收藏

❤ 感谢



Sherry Zhang

关注FinTech喜欢法律的码字猿

10 人赞同了该回答

之前采访过一位智能信贷公司的数据决策总监。他是这么和我说的：大数据风控的核心点在于——**对数据的理解有多深**。做大数据风控的人对数据要极其敏感，因此他们会花很多时间在变量上。

关于把什么变量放在模型里，他和我举过两个例子。

第一个是和电商合作做变量的例子：他们可以通过这些合作看到用户订票的信息、机票的信息，比如公务舱、经济舱这些信息——这本身其实也能说明一个人的基本经济情况。

但是他们会做得更细，会继续做一些叠加或衍生。比如他们会不看公务舱和经济舱的区分，而看飞行每公里的消费单价。因为公务舱和经济舱的价格也会波动很大，有的时候经济舱也有特价票、公务舱也会有优惠活动，所以他们会看每公里的消费金额。

第二个例子是流水话单。他们可以基于同一份电话单，做出很多不一样的变量。比如说用户是否跟某某类的店打过电话？打电话的频次怎么样？趋势怎么样？

如果用户经常跟贷款中介打电话，或者银行催收中心打电话，那用户应该相对比较缺钱，或者是曾经有过违约的历史。

相反，如果用户经常给花店打电话买花，说明用户可能是个“好人”；如果经常给婴儿店打电话，说明他可能有孩子，有孩子的话一般比较稳定、也靠谱一些。

他们会花非常多的时间去衍生这些变量，因为它更直接地反映了这个人的消费行为。当然，也有些时候，这些可能是无用功，有时甚至90%做出来的变量都没有用，但试错筛选出哪怕只有不到10%的可用变量，最终风控效果才是最重要的。

所以，总监觉得，**做风控模型这事儿，一方面是个比较“蓝领”的事儿，因为工作需要做得非常细致。但另一方面，是做模型有时也比较“艺术”，因为这是一个比见仁见智的事情。**

以上。

有兴趣的小伙伴请读全文：zhuanlan.zhihu.com/p/25...

发布于 2017-03-07

▲ 10



💬 4 条评论

➦ 分享

★ 收藏

❤ 感谢



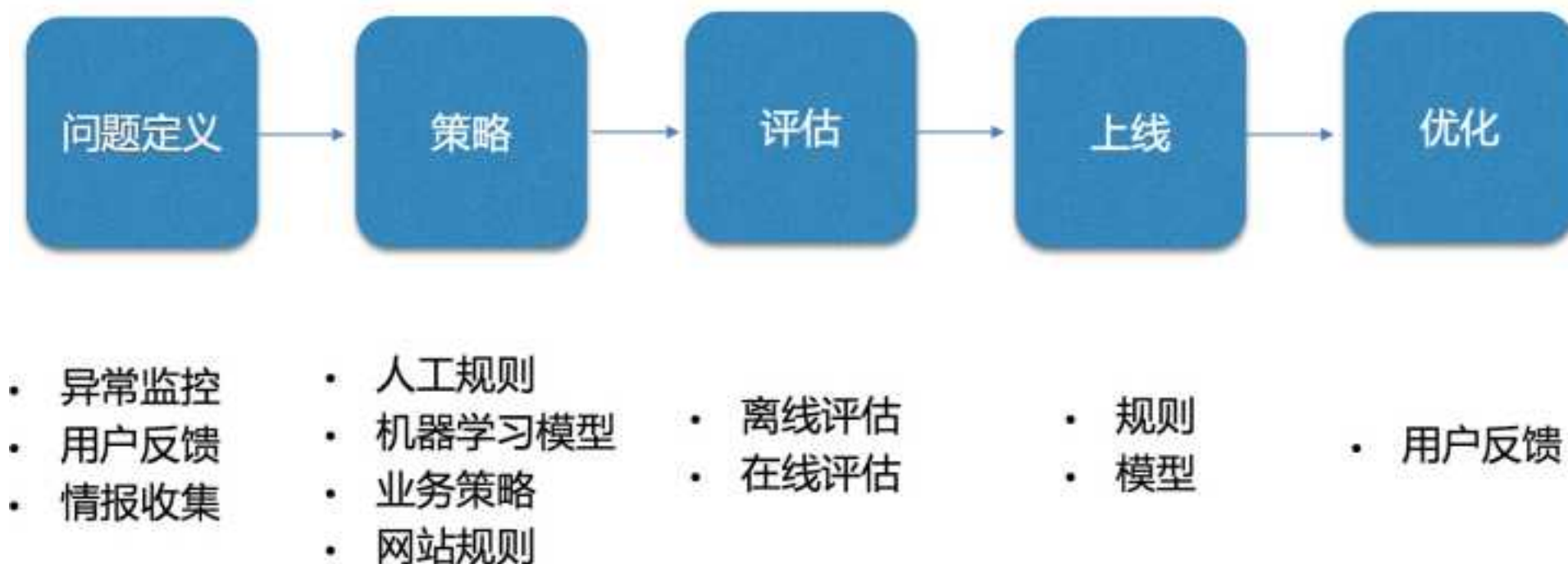
匿名用户

19 人赞同了该回答

风控是什么：

- 顾名思义,风控就是风险控制,最大程度地控制作弊和欺诈的发生,保障网站的正常运营和用户体验
- 风险和作弊行为的发现、识别和处置
- 风控和反作弊是持续的博弈过程,cat-and-mouse game,时效性强,对抗性强
- 三板斧:rules ; models ; strategy

流程包括：



产品体系：



数据：

一般业务数据：用户、商品、交易、点击、浏览、搜索、评价、服务、处罚等

安全业务数据：设备数据（UA、cookie、MAC、Umid、IMEI、IMSI）、位置数据（IP/LBS/GPS）、行为信息、生物信息、其他

算法：

机器学习：分类、聚类、graph算法

异常检测

图像算法：人脸识别、OCR、图像搜索

绝大多数场景使用RF/GBDT+LR/C5.0

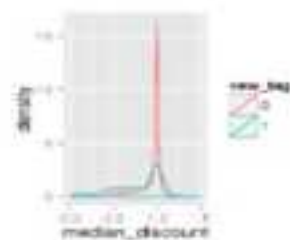
注意点：

- 部分高风险业务,可以投入人力审核,追求更高的准确率/召回率
- 风险(异常)占比少,属于非平衡数据集
- 对抗意识强,模型衰减快,需要结合处置手段
- 风控的成本与回报意识,平衡人力和风险
- 能够采用更复杂的算法,但需要平衡用户体验和可解释性

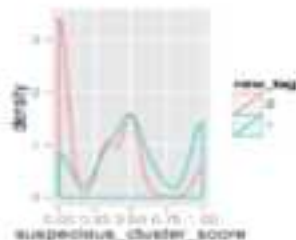
效果评估：

case 1：如何判断某笔交易是否虚假？

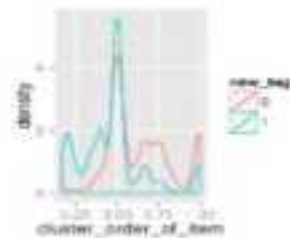
折扣率中位数



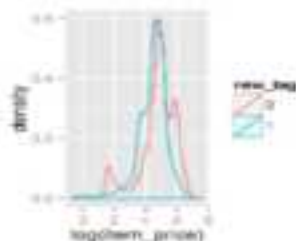
卖家所处聚类的得分



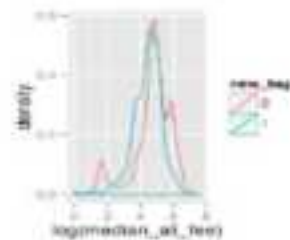
商品价格所处的聚类ID与卖家数量



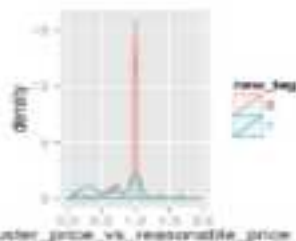
商品价格



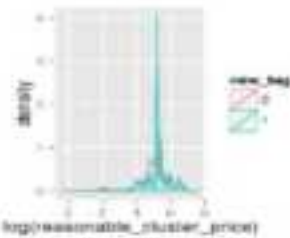
成交价中位数



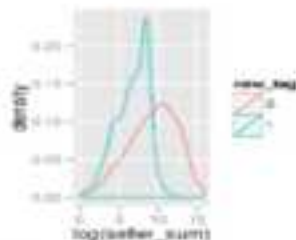
商品所处卖家价格log公允价值

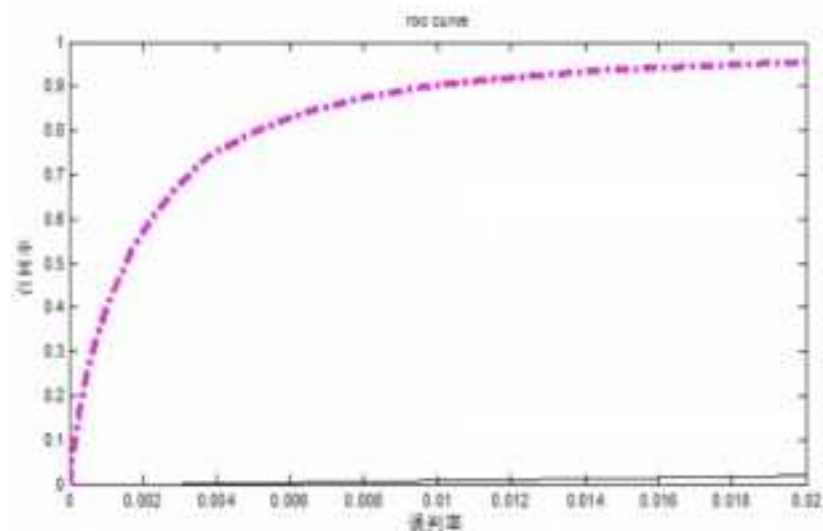


SPU公允价格



卖家评分





Threshold	Class 0			Class 1			PrpMis
	NCorrect	NMisclass	PrpMis	NCorrect	NMisclass	PrpMis	
00000		370	100.00000	159		.00000	69.94329
01000		370	100.00000	159		.00000	69.94329
02000		370	100.00000	159		.00000	69.94329
03000		370	100.00000	159		.00000	69.94329
04000		370	100.00000	159		.00000	69.94329
05000		370	100.00000	159		.00000	69.94329
06000		370	100.00000	159		.00000	69.94329
07000		370	100.00000	159		.00000	69.94329
08000		370	100.00000	159		.00000	69.94329
09000	16	354	95.67568	159		.00000	66.91871
10000	82	288	77.83784	159		.00000	54.44234
11000	139	231	62.43243	156	3	1.88679	44.23440
12000	180	190	51.35135	156	3	1.88679	36.48393
13000	209	161	43.51351	155	4	2.51572	31.19093
14000	231	139	37.56757	153	6	3.77358	27.41021
15000	245	125	33.78378	152	7	4.40252	24.95274
16000	253	117	31.62162	152	7	4.40252	23.44045
17000	260	110	29.72973	152	7	4.40252	22.11720
18000	267	103	27.83784	152	7	4.40252	20.79395
19000	281	89	24.05405	151	8	5.03145	18.33648
20000	287	83	22.43243	150	9	5.66038	17.39130
21000	293	77	20.81081	150	9	5.66038	16.25709
22000	296	74	20.00000	148	11	6.91824	16.06805
23000	301	69	18.64865	148	11	6.91824	15.12287
24000	305	65	17.56757	148	11	6.91824	14.36673

其他：

- 具体的问题定义需要从业务的漏洞、运营规则、法律等方面去思考判断
- 处罚的机制需要平衡用户的体验
- 样本和特征、召回都需要大量领域的知识沉淀

- 评分和识别逻辑要能讲的通

编辑于 2017-06-14

▲ 19



💬 4 条评论

➦ 分享

★ 收藏

♥ 感谢

收起 ^



许铁-巡洋舰科技

微信公众号请关注chaoscruiser ,铁哥个人微信号ironcruiser

1 人赞同了该回答

作者：许铁-巡洋舰科技

链接：[风险管理中的物理思维 - 混沌巡洋舰 - 知乎专栏](#)

来源：知乎

著作权归作者所有。商业转载请联系作者获得授权，非商业转载请注明出处。

反脆弱的世界观

1： 随机之美。

因为宇宙需要它永葆青春。当无常的巨浪袭来，请在内心里欣赏她的优美。

我说复杂性和随机性是美的，因为它其实是宇宙的创造性力量，它摧毁，破坏，并选择强者，让宇宙常常更新，常常进步。就像印度教的那位大神 Shiva，她优美的舞蹈永不停息，这一边毁灭世界，那一边创造新生。人生的优美，在颠覆性和建设性的力量的博弈平衡，发生在你身上的无常，是在赋予你新生。如果没有永恒不停的随机性，我们将生活在一个不值得一活的世界。富者恒富，贫者恒贫。拖勒密的宇宙里，静静的供奉起亚里士多德的神像。一个可以精确预测的世界，出生等于死亡。

2： 世界是非线性的。

复杂系统的运转之所以复杂，源自其组成元素之间的非线性相互作用。在这里， $1+1$ 等于2的机会几乎为零，它大于二或者小于二，**复杂系统的元素之间通过非线性作用关系组成复杂网络，其具有的复杂因果关系，往往非我们穷思竭虑所能达。**

非线性告诉我们什么？规模越大的系统往往越脆弱。

如果你知道墨菲法则，你会更加理解这点。墨菲法则说，如果一个系统可能出错，它终将出错。其实这就是描述基于复杂性产生的脆弱性。当一个过程，一个系统，充满相互关联的步骤或者元素，而过于复杂，它的崩溃几乎是一个时间问题。因为这样的系统由于非线性效应，一个元素的损坏将导致整个系统的损坏。而由于系统内原件过多，出现一个原件损坏是早晚的事情。所以如果可能出错，终将出错。

但是自然里通过进化留下的复杂系统，却往往可以修正因为规模效应导致的脆弱。比如人类的大脑，大脑有10亿级别的神经元数量，其实很多原件是类似的功能，即使丢掉一部分也可以正常运转。有的人被切去半脑却可以正常生活，即是证明。大脑就是一个典型的九头蛇怪，你砍掉它的一部分，它就在另一部分产生类似的功能。

世界的非线性启示我们对自然法则的敬畏，而不是任意的改变复杂系统。

3：世界是个分布函数

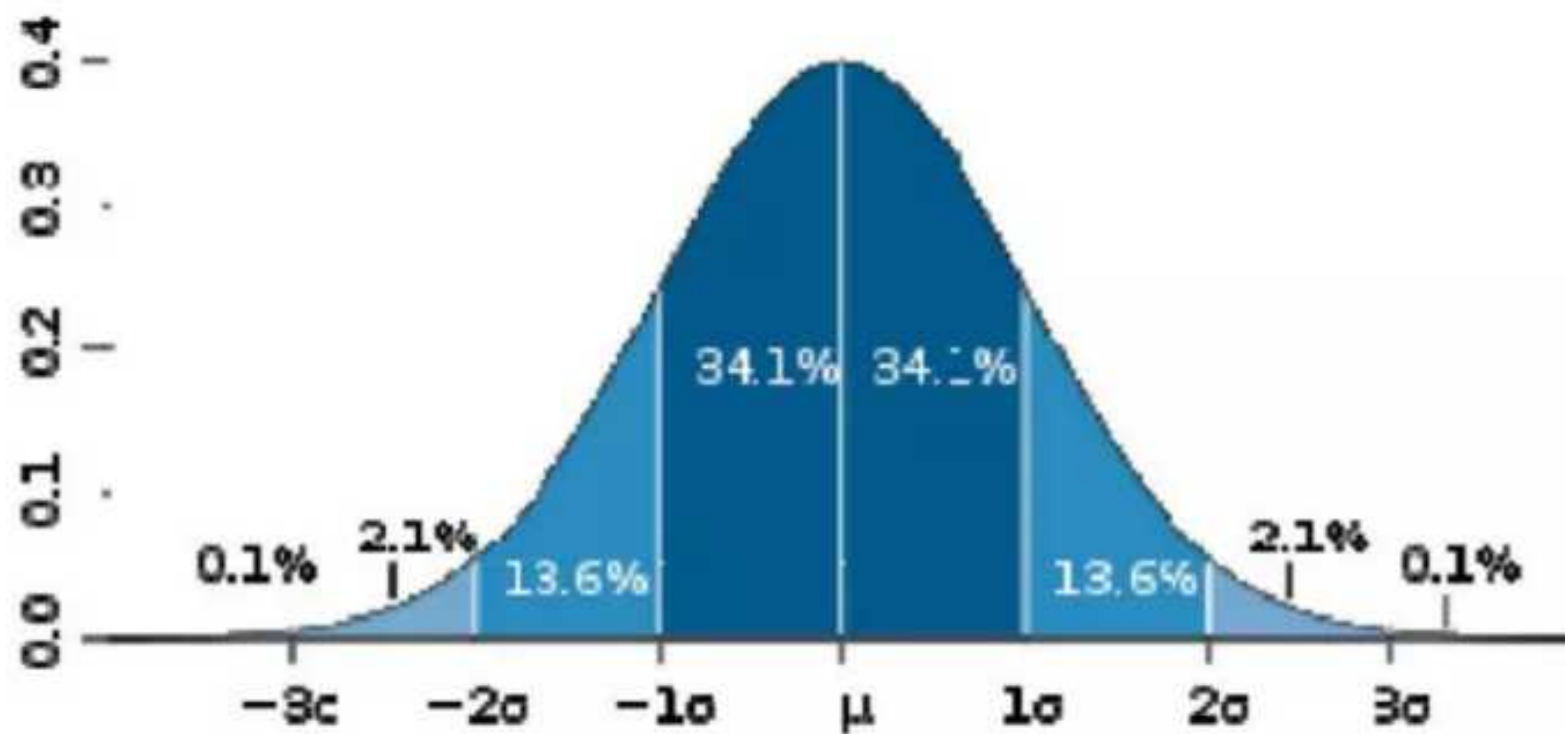
分布函数的世界观告诉我们，看待未来事件，我们要切实的把它看成一个多种可能性的叠加态，而不是非黑即白的确定态。

对于经典物理的系统，平均数往往占支配性作用，热力学里衡量物理属性的各个量，从温度，到压强都是平均数。但是平均数在复杂系统面前，往往不堪一击。其原因在于-分布函数。

分布函数是对随机事件的最佳描述方法，它把一件事的所有可能结果列举出来，并且对应每个结果用一个数表达它发生的可能性。考虑分布，叫我们在高度随机的事情面前考虑各种可能性，并根据每种可能的权重进行决策而非过度倾向某个选项。真正依据分布函数进行思维是很难的一件事，因为我们的大脑的天性是有一些可能无限放大另一些无限缩小，这些往往和我们的心情和刚刚收到的信息有关。比如常见的如果一个新闻刚刚播放了飞机事故，很多人就不敢做飞机，因为心理放大这种事情的概率。

复杂系统的分布函数决定其性质而非平均数，对这个问题我在高斯与天鹅里已经进行了很详尽的描述。**用一句话说，就是幂律函数统治复杂系统。而幂律函数里面极端事件的发生具有比高斯分布大得多的概率。**





Source: Jerry Kemp 2005-02-09 [<http://pbeirne.com/Programming/gaussian.ps>]

混沌巡洋舰

图：幂律的大头和长尾是它的标志，而高斯是大肚。高斯的性质取决于肚子，而幂律则同时决定于大头和长尾。

理解了风险的运作机制，才能更好地实现风险管理。

反脆弱的世界观

1：随机之美。

因为宇宙需要它永葆青春。当无常的巨浪袭来，请在内心里欣赏她的优美。

我说复杂性和随机性是美的，因为它其实是宇宙的创造性力量，它摧毁，破坏，并选择强者，让宇宙常常更新，常常进步。就像印度教的那位大神 Shiva，她优美的舞蹈永不停息，这一边毁灭世界，那一边创造新生。人生的优美，在颠覆性和建设性的力量的博弈平衡，发生在你身上的无常，是在赋予你新生。如果没有永恒不停的随机性，我们将生活在一个不值得一活的世界。富者恒富，贫者恒贫。拖勒密的宇宙里，静静的供奉起亚里士多德的神像。一个可以精确预测的世界，出生等于死亡。

2：世界是非线性的。

复杂系统的运转之所以复杂，源自其组成元素之间的非线性相互作用。在这里， $1+1$ 等于2的机会几乎为零，它大于二或者小于二，**复杂系统的元素之间通过非线性作用关系组成复杂网络，其具有的复杂因果关系**，往往非我们穷思竭虑所能达。

非线性告诉我们什么？规模越大的系统往往越脆弱。

如果你知道墨菲法则，你会更加理解这点。墨菲法则说，如果一个系统可能出错，它终将出错。其实这就是描述基于复杂性产生的脆弱性。当一个过程，一个系统，充满相互关联的步骤或者元素，而过于复杂，它的崩溃几乎是一个时间问题。因为这样的系统由于非线性效应，一个元素的损坏将导致整个系统的损坏。而由于系统内原件过多，出现一个原件损坏是早晚的事情。所以如果可能出错，终将出错。

但是自然里通过进化留下的复杂系统，却往往可以修正因为规模效应导致的脆弱。比如人类的大脑，大脑有10亿级别的神经元数量，其实很多原件是类似的功能，即使丢掉一部分也可以正常运转。有的人被切去半脑却可以正常生活，即是证明。大脑就是一个典型的九头蛇怪，你砍掉它的一部分，它就在另一部分产生类似的功能。

世界的非线性启示我们对自然法则的敬畏，而不是任意的改变复杂系统。

3：世界是个分布函数

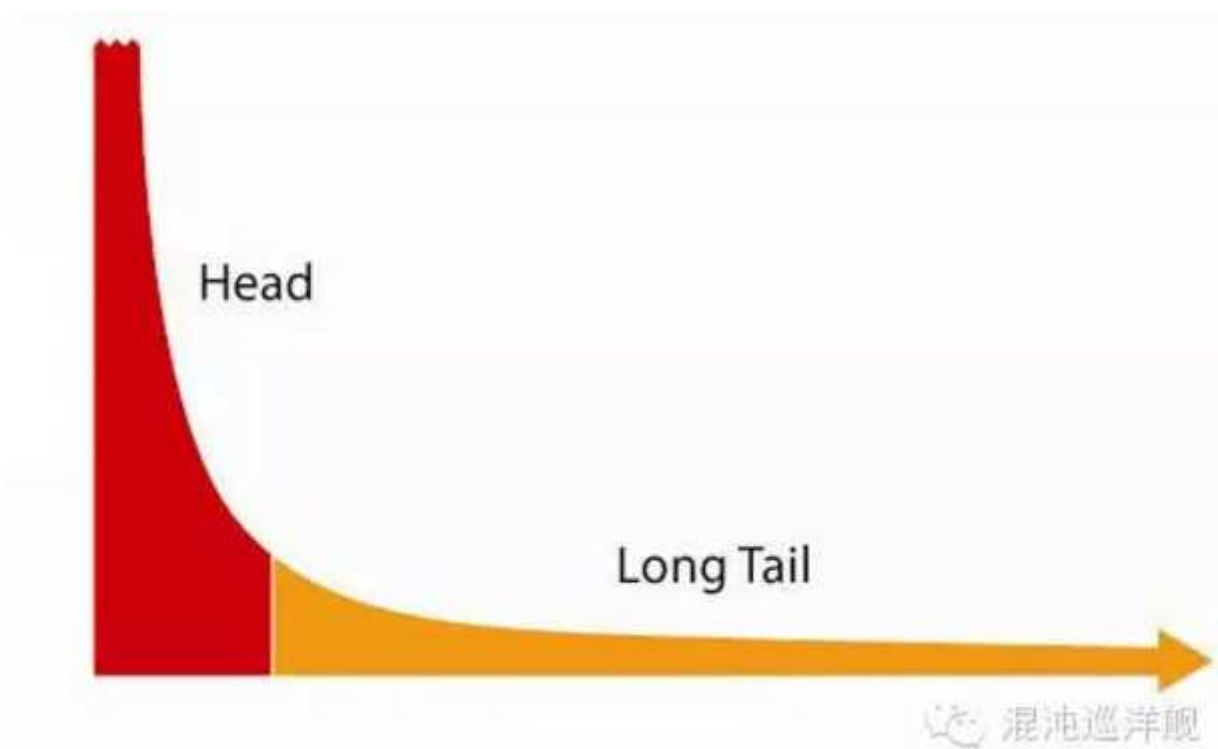
分布函数的世界观告诉我们，看待未来事件，我们要切实的把它看成一个多种可能性的叠加态，而不是非黑即白的确定态。

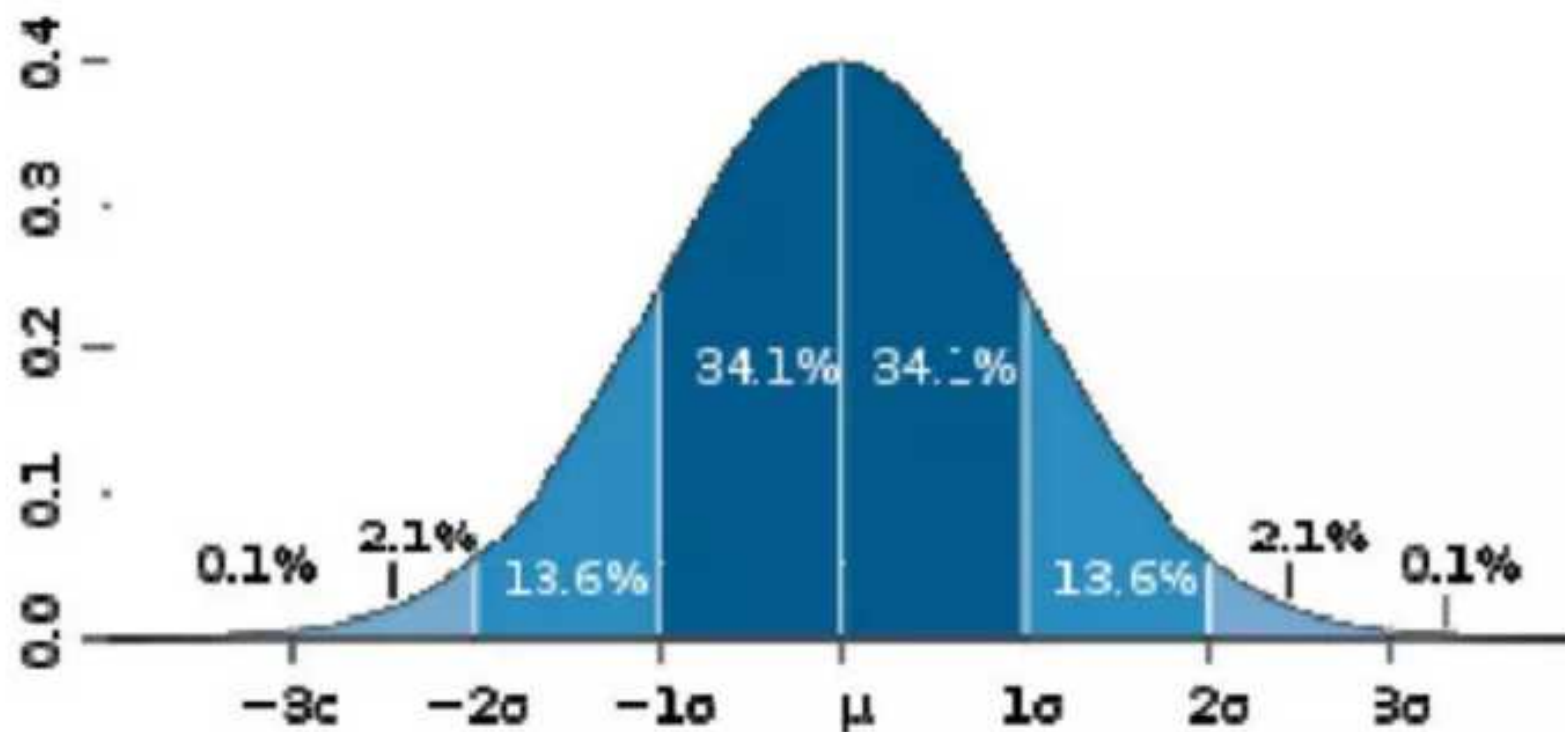
对于经典物理的系统，平均数往往占支配性作用，热力学里衡量物理属性的各个量，从温度，到压强都是平均数。但是平均数在复杂系统面前，往往不堪一击。其原因在于-分布函数。

分布函数是对随机事件的最佳描述方法，它把一件事的所有可能结果列举出来，并且对应每个结果用一个数表达它发生的可能性。考虑分布，叫我们在高度随机的事情面前考虑各种可能性，并根据每种可能的权重进行决策而非过度倾向某个选项。真正依据分布函数进行思维是很难的一件

事，因为我们的大脑的天性是把这些可能无限放大另一些无限缩小，这些往往和我们的心情和刚刚收到的信息有关。比如常见的如果一个新闻刚刚播放了飞机事故，很多人就不敢做飞机，因为心理放大这种事情的概率。

复杂系统的分布函数决定其性质而非平均数，对这个问题我在高斯与天鹅里已经进行了很详尽的描述。**用一句话说，就是幂律函数统治复杂系统。而幂律函数里面极端事件的发生具有比高斯分布大得多的概率。**





Source: Jerry Kemp 2005-02-09 [<http://pbeirne.com/Programming/gaussian.ps>]

混沌巡洋舰

图：幂律的大头和长尾是它的标志，而高斯是大肚。高斯的性质取决于肚子，而幂律则同时决定于大头和长尾。

理解了风险的运作机制，才能更好地实现风险管理。

发布于 2016-10-09

▲ 1

▼

添加评论

分享

★ 收藏

♥ 感谢

收起 ^



匿名用户

8 人赞同了该回答

风控流程：目标 + 过程 + 结果

风控目标需要考虑到风控目标的数据量，数据类型，数据误判代价，数据的更新频率等。

风控过程中，

数据量大的时候可能数据算法或者机器学习方法更合适，但是数据量或者坏样本数据较少的情况下类似判断方法或者权重判断更加优秀。数据类型也影响着算法的选择，不是所有数据都适合市上主流算法，比如过多的名义变量下，lg模型并不是很合适，需要预处理或者选择其它方式。至于误判代价，是犯错成本，犯错成本更大的情况下，有时候算法结果需要规则修正，并不是说从始而终的算法解决。最后频率上考虑上线后的数据及时性，是考虑固定规则还是实时的动态判断。之外，还有很多很多因素。

风控结果上，考虑到反馈形势，是单纯的0/1还是0-1，需要考虑犯错成本。

至于算法选择上，行业比较好的算大家都在用的逻辑回归，也比较常见的是gbdt，但是算法是死的，可以考虑不通过的损失判断函数之类的，个人用过的其它比较好的还有：

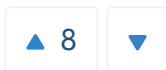
SVM 作为史上最强分类器，解决这种小样本复杂问题的利器，综合量化判断模型确实有不少采用SVM的，不过不一定要做直接判断结果，可以做 backup - key。

DNN和CNN在部分情况下效果很不错，大流量，不计较小损失的情况下，效果可以复制。个人感觉，用的场景很苛刻，也不好解释，但是有小场景下单意料之外。

至于规则修正选择上，很多key - value的排序方法，或者说一些传统的ahp方法等等，在数据量缺失等前期算是很不错的决策方案。

其实，个人在敏感部门，只能匿名，个人想法，欢迎讨论，拒绝水表。

编辑于 2016-08-21



3 条评论

分享

收藏

感谢



Magic

专注银行数据挖掘与人工智能，欢迎交流！

2 人赞同了该回答

首先，貌似美国基本上都用三大征信局的信息，最传统的评分基本上都是用FICO来做的。同时，各家平台也会尝试着用机器学习、神经网络等大数据处理方法，但探索的居多，有结果的少。

其次，现在国内互联网金融基本的业务规则并没有太大变化，大数据整体应用也只是刚刚开始（不是做一两个分析案例那种），所以fico评分的框架思路会持续沿用，尝试加入新的变量看看模型效果。

最后，还是看好大数据的发展，看好区块链技术的发展，期待技术逆袭。

发布于 2016-02-15

▲ 2



💬 添加评论

🔗 分享

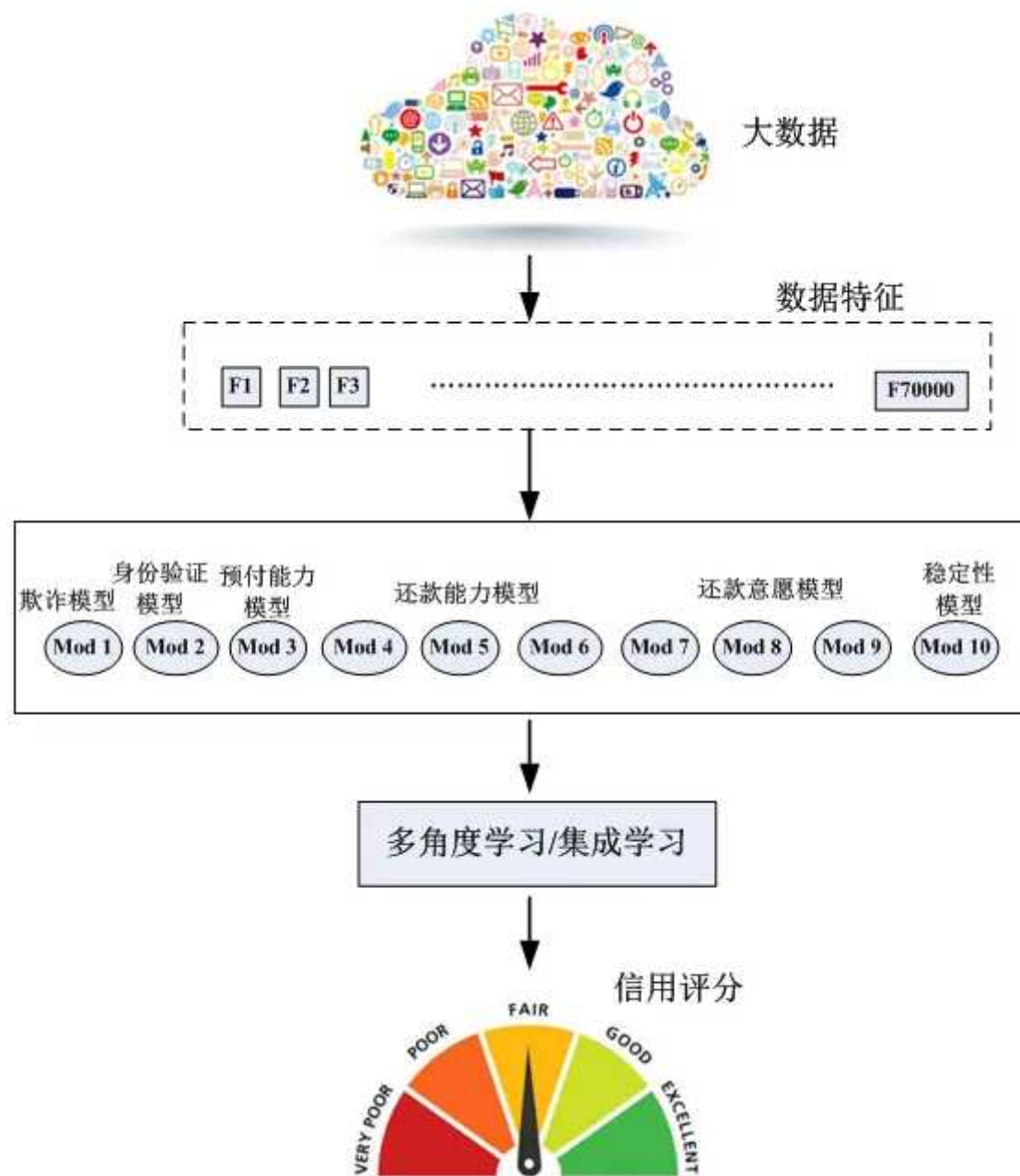
★ 收藏

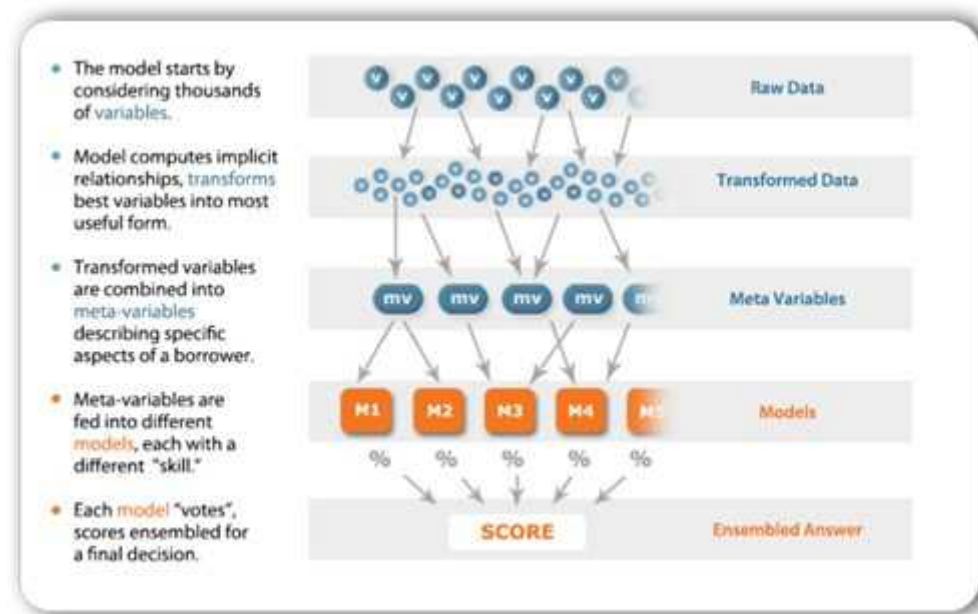
💗 感谢



Sherrie雪小梨

20 人赞同了该回答





上面两张图片都是关于著名的互联网金融公司Zest Finance。

关于有效性，它的官网是这么写的：

ZestFinance underwriting models offer a 40% improvement over the current, best-in-class industry score
也就是比业内风控平均水平高出40%。

编辑于 2015-11-19

▲ 20



💬 2 条评论

➦ 分享

★ 收藏

❤ 感谢

收起 ^



铭铭

互联网金融民工

1 人赞同了该回答

现在国内的云图征信是专门做这一块儿的，贷中、贷前、贷后有14种大数据风控模型，实时动态监控，可以参考 yuntucredit.com

发布于 2016-09-07

▲ 1 ▼ 添加评论 分享 收藏 感谢



李Stine

观察新鲜业态并恶补财务知识.....Matlab Python苦手

8 人赞同了该回答

不邀自来；觉得光摆一些Fancy的图表流程一点用都没有...骗几个外行赞而已...

我就只谈一谈保险行业的欺诈风控吧...

目前保险行业做的风控最好的是平安。其他几个基本停留在业务风控上。为什么这么说呢？因为平安是少有的做了业务数据外的操作数据风控的集团。

其次人保人寿也都搭建了基于业务数据的风控模型，构建了黑名单灰名单，有一套完整的风控体系。（主要说的是寿险财险车险）单就结果来看，事后查出并追回的骗保费用全国来看少说也有几千万级别了。

我倒觉得现行模型表现并不如树模型好，尤其是保险行业树模型反而更贴近业务更能Make Sense地发现风险点。

我们也尝试做了预测总模型，效果怎么说呢，比猜好很多，但是依旧离实用比较少。

另外银行的欺诈方面建行，汇丰都做得不错，不过细说就深了，因为银行内部风险特别大，此外欺诈风险只是银行面临风险很小的一部分。

总之数据质量很重要。比它更重要的是领导的支持，大数据风控这个东西你没个领导的支持，没个管理的部门，没有激励的机制，只有模型，都是空的。

发布于 2015-12-06

▲ 8 ▼ 7 条评论 分享 收藏 感谢



神经病患者

神经病患者

4 人赞同了该回答

手机答题，忽略排版！帮题主缩小一点范围，我姑且把题主的模型认为是现在的互金公司的风控模型，而不是一般的互联网公司防盗号、防薅羊毛的那种。

现在的这一类风控模型大多仍然是沿用之前传统银行信用卡中心那一套，俗称评分卡。这种评分卡一般分为三种，分别是a卡，用于客户申请信用评估；一种是b卡，也被称为行为评分卡，用于评估客户贷中的风险；一种是c卡，用于催收策略。前两种模型有过接触，最后一种不太清楚具体的业务方法。

现在一般说的风险模型就是a卡，也是三种模型里面最重要的，因为如果在前面能成功的把坏客户挡在申请外面，后面两种模型就显得无足轻重了。相比之前的信用卡时代，现在的小贷公司能拿到的数据远远比以前拿到的数据要多，但触碰到的隐私红线的机会也会更多，这其中也和现在第三方征信公司的野蛮发展有很大关系。通常在申请的时候，客户会被要求填几个基本信息，如姓名，身份证，手机号及其他一些人口属性信息，贷款公司拿到几要素之后就会去抛第三方征信公司的数据接口，从而拿到自己没有能力拿到的数据。综合各方的数据，一般拿到的数据可以分为这么几类：一类是人口属性信息，这其中比较重要的如性别，年龄，学历，行业等；一类是device信息，包括手机型号，ip地址，lbs地址等；一类是借贷信息，如在各种机构下面的申请，借款，还款信息等；还有一类是补充信息，这类信息通常会触碰隐私红线，如话单信息，通讯录信息，app使用信息，历史lbs轨迹等。另外多一嘴，很多人认为话单的作用会随着微信的盛行而大幅减弱，从而完全失效，但是在实际使用中，尤其是在负面信息的表达上，话单的效果还是相当抢眼的，另外一个比较好用的信息就是设备上app的安装和使用情况，简直就是弥补了多头借贷的信息，并且相比借贷信息，也增加了一部分正面信息，这能更好的提高模型区分好坏的能力。

现在风控模型中最常用的算法仍然是逻辑回归，它的地位这么稳也是有一定道理的，最大的好处就是可解释性，可解释性在这一领域有很大用处，这一好处不仅仅是对客户好解释，这也和现在小贷公司获取外部数据有关，现在公司从外部数据拿到的数据一般不是底层的原始数据，而是中间加工过的数据，有加工就意味着不稳定性，这种不稳定也会造成风控模型的不稳定，所以当某一变量发生较大变化时，如何评估对现有模型的影响，比如预测的结果会前偏还是会后偏，如何调整策略等。其他的算法，如随机森林，gbdt，xgboost等都会做一些尝试。

模型建好之后，会面临比较复杂的测算过程，如模型在外推样本上的稳定性、有效性如何，如何定策略，在这一策略下，我的拒件率，逾期率会怎么变化，都需要评估出来

所以在我的认知当中，算法和跑模型不是最重要的，预测目的，建模样本选取，变量衍生，模型测算，策略制定才是关键。

后续有空再针对某一点做详细的介绍。

编辑于 2017-08-20

▲ 4



● 添加评论

➤ 分享

★ 收藏

♥ 感谢



iamcodylee

Life is ephemeral, live it up

3 人赞同了该回答

用什么模型不是最重要的吧，难在feature选择上，不同的feature出来的结果可能会有很大的差异。

发布于 2016-08-25



3 条评论

分享

★ 收藏

♥ 感谢



知乎用户

公众号：逍遥吃玩——淡江湖玩玩，懒侠客吃吃

3 人赞同了该回答

这话题开放的.....

风控分贷前风控，贷中风控，贷后风控，说穿了，就是别人找你借钱，前期你要评估这钱借出去有没得还，该以多少点借出去，借出去之后到钱还回来整个过程你还要盯着，避免他突然跑路或者是

风控具体指的是什么？万一他耍赖皮说不还你要怎么让他还。

数据么，当然是金融行为数据最相关了，现在用电商啊社交什么的数据来做分析，就跟你走访时候了解他朋友他平时都是些什么生活习惯差不多啦。算法每家各不同，但是应该没有谁真的愿意讲出来吧，毕竟是核心。

风控应用案例：蚂蚁花呗借呗、腾讯的微粒贷都算是啊。

至于效果么，估计怎么也得明年才会有相关的报告出来，毕竟数据积淀需要时间，而且前期应该不会太好看。然而大数据风控是趋势，各家产品在市场的不断调整中竞争优化，会越来越有效的。前期尝试或者作为必需信息来源的补充渠道嘛。产品那么多肯定有个自然的筛选过程的。蜜蜂数据对接多家大数据征信产品，也是会经过筛选对比的。虽然目前发展的时间跟整个行业一样，不长，但是可以为征信公司和平台的沟通搭桥，避免他们相互交叉——对接试错来着！

发布于 2016-04-13



3 条评论

分享

★ 收藏

♥ 感谢



王乐石

律师

3 人赞同了该回答

感觉大数据对风控没有明显的作用。针对消费者的小贷业务也是根据存款额，消费额来定，同样是万分之五的利率。逻辑和银行是一样一样的。而且借贷到期后强制扣款，这点比不上银行。不看好。总体上，国内对大数据抱有过高的不切实际的期待。实际上消费者的消费决策并不能依靠过去的数据来进行判断。it做流程规范的空间，做娱乐的空间也是有限的。在信用社会，总会有一只或者无数只猪在天上飞，问题是它们掉下来时别砸到自己。

发布于 2015-12-04

▲ 3 ▼ 2 条评论 分享 ★ 收藏 ♥ 感谢



耿树文

关注互联网金融产品、运营、风控

2 人赞同了该回答

楼上有很说了很多技术方面的，这些技术的运用挖掘了用户的还款能力、还款意愿。

但还有一个核心问题，不是用了什么模型，而且是风控的有效性如何评价？

风控的有效性，一定是定价、准入、坏账达到均衡。

不是简单说好坏的二元法则，那基本上就是传统风控，传统风控就是一些准入条件，条件一旦达不到就拒，这跟互联网大数据的原则是相违背的，互联网大数据风控要实现的是“千人千面”。

发布于 2017-03-22

▲ 2 ▼ 添加评论 分享 ★ 收藏 ♥ 感谢



刘彧

Data

2 人赞同了该回答

你有什么大数据先

发布于 2015-11-18

▲ 2 ▼

添加评论

分享

★ 收藏

♥ 感谢



匿名用户

1 人赞同了该回答

可以看下我们做过的一个案例：

数据风控的探索实践,机器学习识别欺诈

图文原文：mp.weixin.qq.com/s/cXGR...

专栏：[拓端数据部落](#)



在信息爆炸时代，“信用”已成为越来越重要的无形财产。”数据风控“的实际意义是用DT（Data Technology）识别欺诈，将欺诈防患于未然，然后净化信用体系。

挑战

信贷风险和欺诈风险是消费金融业务发展中最重要两种风险，信息不对称是导致这些风险的主要原因。

“数据防欺诈”是数据风控武器之一。这种武器的力量的重要保证是数据和信息收集的完整性和准确性。通过这些有价值的数据，找到欺诈者留下的线索，以防止发生欺诈。

实施过程

｜ 用户立体化呈现——多维数据采集

tecdat深入分析用户的基本属性、社会属性、消费者行为、兴趣偏好、社会偏好、资产特征、信用特征等数据，通过数据挖掘，使用户更加立体化地实时呈现。

｜ 挖掘潜在的团伙欺诈——社区发现算法

一方面，基于机构的存量数据，运营商等数据构建复杂的网络。同时，采用社区挖掘算法实现风险分组。在此基础上，我们训练机器学习模型。

｜ 建模的原材料 —— 特征工程

建模的第一步是特征工程，众所周知，特征是机器学习建模的原材料，对最终模型的影响至关重要。数据和特征比模型更重要，数据和特征决定了机器学习的上限，而模型和算法逼近这个上限。特征加工和衍生工作越完备，那么构建的机器学习模型效果越好。**但是，面对不同数据，不同业务场景，特征加工衍生往往是最耗时间与资源的工作。**

尤其在弱数据方面，充斥着大量文本、时序类数据，人工特征定义的方法天然存在较大局限性。

tecdat引入基于机器学习的特征提取框架（如 random forest，SVM，CNN）来适应不同的数据类型，自动从大量复杂的非结构化数据中产生高质量的特征，**完成模型训练后可以输出特征的重要性**，结合多种方法进行特征选择和解释。

｜ 和而不同——集成模型

具体的模型，我们知道在弱势数据的基础上加工和衍生的特点，机构往往面临很多特征维度，从数千到数万以上，非常稀疏。超出了传统风控的基于评分卡系统的建模能力。

tecdat引入集成模型(ensemble models)来解决这个问题。**集成模型从“投票”的思想简单的理解，也就是我们对不同类型的数据使用最合适的子模型（Logistic回归，GBDT，CNN，xgboost），然后每个子模式投票作出决策。**

能够使整体模型的准确度和防止过拟合的能力达到协调，从而达到在总体上的最佳准确度。

复杂的集成模式框架除了当前场景和业务建模具有很好的表现，其另一个重要价值在于可以快速应用于新业务应用，对“冷启动”阶段有非常重要的作用。

结果/效果总结

最后，在线上信用贷场景实践下来，经过多批次多个跨时间段的验证，可以看到，效果上还是有非常直接的提升，模型性能相比传统模型提升了大约30%。

发布于 2017-07-27



添加评论



分享



收藏



感谢

收起 ^



1二叁

小数据分析师

1 人赞同了该回答

我看大家大都是说信贷类，我来简单说下非信贷类的，抛砖引玉。相比于信贷类来说，非信贷类场景复杂且多（薅羊毛，虚假点评，众包作弊等等），label不好标。在信贷类中，至少在事后根据还款行为可以打上标签，而在非信贷类中可能一条事件永远也打不上可靠的标签。

在之前我们团队主要用的是xgboost,LR,RF等模型，也有想过委员会投票法，当然最终结果还要等时间来检验。其中xgboost的优点很明显，效果好，训练速度快。当然也不是所有数据都可以直接丢xgboost（不然的话还要我们干嘛），之前我就碰到过一次，特征工程没做好，导致xgboost认为重要的指标之间有很高的相关性，这样看上去模型效果很好，实际上是没用的。

现在我主要的研究方向是半监督学习，这也是因为可靠的label不好找。因此考虑到是否可以用半监督算法。

最后，第一位是数据，第二位是特征，模型再次之。

编辑于 2017-06-16

▲ 1 ▼ 2 条评论 分享 收藏 感谢



巩开学

国内金融、国际市场开发

1 人赞同了该回答

我也在帮朋友公司找一位风控建模和数据开发方便的专家工程师，看了大家的回复觉得大神不少呢，有意者私聊

发布于 2016-12-05

▲ 1 ▼ 1 条评论 分享 收藏 感谢



知乎用户

知乎

1 人赞同了该回答

1、基于某类特定目标人群、特定行业、商圈等做风控。由于针对特定人员、行业、商圈等垂直目标做深耕，较为容易建对应的风险点及风控策略。

2、基于自有平台身份数据、历史交易数据、支付数据、信用数据、行为数据、黑名单/白名单等数据做风控。

3、基于第三方平台服务及数据做风控 互联网征信平台（非人行征信）、行业联盟共享数据（例如小贷联盟、P2P联盟）FICO服务、Retail Decisions(ReD)、Maxmind服务。

4、基于传统行业数据做风控 人行征信、工商、税务、房管、法院、公安、金融机构、车管所、电信、公共事业（水电煤）等传统行业数据。

5、线下实地尽职调查数据

包括自建风控团队做线下尽职调查模式以及与小贷公司、典当、第三方信用管理公司等传统线下企业合作做风控的模式。线下风控数据也是大数据风控的重要数据来源和手段。

发布于 2016-09-07

▲ 1 ▼ 添加评论 分享 收藏 感谢



大罗森

简简单单，个人订阅号：有时回廊

1 人赞同了该回答

排除银行传统的个人信贷而言，大数据风控在互联网金融领域目前实用价值太低。

中国是熟人社会，在一线城市，个人迁移成本太低，在二三线城市个人的信贷需求又不够高，这是最难克服的问题。

个人信贷大数据收集，如果是利用互联网社交圈信息收集（个别P2P平台就是这么做的），存在两大问题。

一是个人信息的真实情况存疑，包括个人的社交圈、财力评估、还款意愿评估。

二是即使有平台或者第三方构建了相对科学的评估模型，但缺少政府对全体公民征信的约束效用，个人违背这个数据模型的代价也是极有限的。

脱离政府背书构建的任何大数据风控都是概念、自欺欺人。

发布于 2016-03-18

▲ 1 ▼ 3 条评论 分享 收藏 感谢



Jessica

关注DATA的小白

1 人赞同了该回答

模型构建后 影响因子需要不断调整 完善 然后才能用于应用

发布于 2015-12-03

▲ 1 ▼ 添加评论 分享 收藏 感谢



景云

单身狗，程序猿，技术宅

1 人赞同了该回答

关键是数据和特征，模型都大同小异

发布于 2015-11-20



💬 1 条评论

✈️ 分享

★ 收藏

♥️ 感谢



天创信用

国内领先的大数据风控服务商 (tcredit.com)

想了解的话欢迎来看看哦~

66号学苑出品

风控大牛手把手教你 搭建企业级信用评分模型

12月14日 (周四) 20:00

「乔 杨」



ZRobot CEO
前美国发现金融
资深风控专家

扫描二维码

即可进入课程页面



#独家课程#12月14日（本周四 20:00），66号学苑携手ZRobot CEO乔杨开设信用评分模型系列课程，从概念应用、数据基础、数据挖掘技术、开发流程、实战案例等方面，手把手教你如何搭建企业级信用评分模型。课程地址：[干聊](#)

编辑于 2017-12-12

 0   添加评论  分享  收藏  感谢

收起 ^



qsdemm

4_

发布于 2017-04-21

 0   添加评论  分享  收藏  感谢



lynnrichie

涉及到欺诈这块社交网络，大家是如何实现社交图谱反欺诈的呢

发布于 2016-11-18

 0   1 条评论  分享  收藏  感谢



萧鸣

只能对量化信用风险做一些历史数据的参考。别忘了信用风险只是全面风险的一小部分

发布于 2015-11-20

 0   添加评论  分享  收藏  感谢



达达的马蹄

过客...

我国信用体系并不完善，风控还得靠人

发布于 2015-11-18



1 条评论

分享

★ 收藏

♥ 感谢



讨厌红楼梦

乡村金融民工

个人以为可以尝试，不可以依赖。

发布于 2015-11-17



添加评论

分享

★ 收藏

♥ 感谢



看看而已

踮起脚尖，高一点是一点

正准备找一个大数据风控建模数据分析负责人，有意向的朋友可私下聊哟。

编辑于 2016-11-23



3 条评论

分享

★ 收藏

♥ 感谢



Wei Zhang

安全存储工程师 微信公众号：gsgsoft

各位前辈讲的真好。。。。

发布于 2016-02-28



添加评论

分享

★ 收藏

♥ 感谢

 写回答

5 个回答被折叠（为什么？）