A Model Framework to Estimate the Fraud Probability of Acquiring Merchants

by

Ye Zhou

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Business Administration

Approved March 2015 by the
Graduate Supervisory Committee:

Hong Chen, Co-Chair
Bin Gu, Co-Chair
Xiuli Chao

ARIZONA STATE UNIVERSITY

May 2015

基于历史数据的银行卡收单商户欺诈风险预测模型

作者：周晔

工商管理博士
学位论文

研究生管理委员会
批准于二零一五年三月：

陈宏，联席主席
顾彬，联席主席
赵修利

美国亚利桑那州立大学

2015 年 5 月

ABSTRACT

Using historical data from the third-party payment acquiring industry, I develop a statistical model to predict the probability of fraudulent transactions by the merchants. The model consists of two levels of analysis – the first focuses on fraud detection at the store level, and the second focuses on fraud detection at the merchant level by aggregating store level data to the merchant level for merchants with multiple stores. My purpose is to put the model into business operations, helping to identify fraudulent merchants at the time of transactions and thus mitigate the risk exposure of the payment acquiring businesses. The model developed in this study is distinct from existing fraud detection models in three important aspects. First, it predicts the probability of fraud at the merchant level, as opposed to at the transaction level or by the cardholders. Second, it is developed by applying machine learning algorithms and logistical regressions to all the transaction level and merchant level variables collected from real business operations, rather than relying on the experiences and analytical abilities of business experts as in the development of traditional expert systems. Third, instead of using a small sample, I develop and test the model using a huge sample that consists of over 600,000 merchants and 10 million transactions per month. I conclude this study with a discussion of the model's possible applications in practice as well as its implications for future research.

摘要

本课题的目的是针对银行卡收单欺诈行为，根据多元历史数据，建立网点级别的欺诈概率预测模型，并进一步汇总至商户级别，从而获得商户收单欺诈概率预测模型。该模型将最终投入实际的商业运营，缓释收单业务的欺诈风险。本课题将试图在三个方面有所创新：一是针对商户层次建立预测模型，而不是以交易或持卡人为对象，从而避免交易层次模型引起的处理效率下降以及客户体验不佳等问题；二是运用机器学习中的数据分析方法，把所有特征变量纳入学习范围，并由概率型非线性回归进行二元分类，而不是传统的、基于行为分析经验的专家系统；三是数据处理直接针对海量原始数据操作，而不是先小样本建模，然后再验证假设的传统方法。最后，对课题建立的数据可能产生的前景作了展望。

鸣谢

衷心感谢上海交通大学上海高级金融学院教授陈宏、赵修利、顾彬给予的指导，也感谢汇付天下同事赵敏博士、李莉小姐提供的无私帮助。

目录

页码数

章节 页码数

章节 页码数

表格目录

表格 页码数

图表目录

图表 页码数

章节 1 绪论

1.1 课题研究的目的和意义

自有支付以来，支付服务商一直在平衡支付工具的便捷性和风险，一方面尽可能地实现支付手段的便捷性，使付款方和收款方在时间和成本上获益；另一方面又要防范各类信用和欺诈风险，使付款方、收款方敢于使用支付产品和服务，支付服务商亦能生存、盈利。

银行卡自 1950 年代发明以来，风靡全球，迅速成为现代社会的主流支付工具之一。截至 2013 年底，全球已发行信用卡、借记卡达 112 亿张（来源 RBR 零售银行研究）。中国银行卡产业也是突飞猛进。自 1986 年发行第一张银行卡以来，至 2013 年底，已拥有共 43 亿张各类银行卡，商户数已逾千万，支付场景已涵盖商业零售、批发、对公结算、投资等多个领域，支付形式也包括线下收单、互联网支付、移动支付、电话支付等等。

但是，银行卡给付款和收款带来巨大便利的同时，也伴随了巨大的风险。支付行业一直以来都面临了日新月异的欺诈，如伪卡、盗卡、套现、钓鱼、侧录、抵赖、移机等等，给持卡人、发卡机构、商户、收单机构和银行卡组织等带来了极大的麻烦，甚至威胁。随着发卡量的增加、IT 技术的普及、全球化程度的提高，欺诈越来越向团伙化、跨国化、网络化发展，导致了犯罪手段更新快、传播快、波及面大，欺诈金额也逐年攀升。

中国目前收单业务正处在一个产业转折点。伴随着发卡的普及和通信网络的覆盖，收单业务正从大中型商户走向小微商户，正从中心城市走向二、三线城市，乃至广大乡镇。小

微商户由于规模小、经营多变、工商注册信息不全、经营者法律观念淡薄等原因，容易成为风险薄弱环节，因此，在这个收单业务转变之中，银行卡的欺诈也从以卡欺诈为主，转向了商户欺诈或持卡人与商户合谋欺诈，以不当手段从商业银行获取资金，给发卡机构和收单机构带来了巨额风险损失。

尤其值得一提的是，在中国当今特定的金融体制之下，小微商户从商业银行正规获取融资的可能性极低，同时，在融资渠道中，商业银行贷款成本与民间借贷存在着很大的价格差异。这两个因素相加，导致了大量小微商户的套现行为，即利用 POS 收单或网络支付从商业银行的信用卡中获取免息或低价资金。

因此，建立对商户欺诈行为、尤其是对小微商户欺诈行为的风险监控，有预见地识别商户的欺诈行为，成为支付行业风险控制的重要任务。它的意义在于：

1) 降低收单机构、发卡机构的风险损失，尤其避免消费者的损失；

2) 建立支付行业和消费者对小微商户受理银行卡的信心，从而使得收单业务可以推广至千万数量的小微商户，让占绝大多数的、诚实经营的小微商户也能从现代支付工具中受益。

本课题的目的就是建立商户的欺诈预测模型。我们从近 60 万遍布全国（除青海、西藏）的小微商户数据入手，通过对静态信息、交易数据、外部风险数据的清理、分析，运用

机器学习中的信息值提取、聚类分析和 Logistic 回归方法实现分类算法，最终建立可以投入商业运营的商户欺诈概率模型。

1.2 支付系统的参与者及支付流程简介

由于本课题只涉及线下收单业务，故不对互联网支付、移动支付和 O2O 做详细阐述。下面的简介主要针对线下收单业务。

自上世纪 50 年代发明信用卡以来，经过半个多世纪，银行卡产业格局基本稳固，普遍形成了以发卡机构、收单机构、银行卡组织、持卡人和商户五个群体构成的产业链。中国自 1986 年中国银行发行第一张银行卡以后，各大银行分别建立了独自的发卡和收单系统。从 1994 年起，经过近八年的金卡工程建设，终于在 2002 年成立了中国银联，初步实现“一卡在手、走遍神州”的愿望。2011 年央行开始发放支付牌照，准许非金融机构进入收单领域，2013 年央行颁布新版《银行卡收单管理办法》。至此，产业链框架基本清晰，与国际通行的业务框架大致相似，采用的业务和技术标准也基本一致。

1) 发卡机构

发卡机构（Card Issuer）就是向消费者（少量对公）发行信用卡或借记卡的机构，是典型的 B2C 业务。与美国体制不一样的是，在中国现行体制之下，发卡机构必须具有商业银行资质，业务必须经过银监会的报备，不得跨地发行银行卡。目前所有发卡程序规定，初始

签约、身份核实必须面对面进行。目前中国全国性发行信用卡的银行已经超过 20 家，共发行约 10 亿张信用卡（也称贷记卡），200 余家银行发行借记卡，绝大部分借记卡和信用卡使用银联品牌，但也有部分银行与 Visa、MasterCard、American Express (AE) 等发行双币种卡或国际卡。

2) 收单机构

收单机构（Merchant Acquirer）是帮助商户（包含政府部门、企事业单位）接受、处理银行卡支付的机构，承担包括商户签约、终端布放、交易受理、路由转发、资金结算等任务，是典型的 B2B 业务。中国在 2003 年之前，几乎所有的收单业务都由商业银行主导，彼时银行既是收单行又是发卡行。2003 年中国银联成立银联商务服务有限公司（简称银联商务）以后，国内近一半的收单业务转由银联商务经营。2011 年开始，央行陆续向非金支付机构发放约 45 张收单牌照，中国收单机构组成遂变得丰富多彩，竞争逐步加剧，销售方式也由原来的单一直销模式转变为直销、渠道等多种方式。

3) 卡组织

理论上，卡组织（Card Association）是由发卡机构和收单机构共同发起成立的非营利会员机构，目的是建立和维护统一的银行卡品牌；通过会员共同协商，制定和维持业务规则和技术标准；构建由发卡机构和收单机构等会员共同组成的联盟网络。它的基本任务包括发卡机构与收单机构之间的交易转接、资金清结算、交易争议处理、品牌推广、会员发展与

管理、新产品开发、相关法律推动等等。从国际范围来看，Visa、MasterCard 是最早、也是最领先的卡组织；中国目前仅有的卡组织是中国银联（CUP）。

4) 持卡人

持卡人（Card Holder）是指持有银行卡（借记卡或贷记卡）的消费者。他们与银行签订协议，获取银行卡，可以依照卡组织品牌在不同的终端进行消费、取现、转账、查询等操作。持卡人获取银行卡的目的是能享受世界范围的、线上线下支付的便捷性；同时其权益也受法规严格保护，如美国针对信用卡持卡人的《Z 法案》和针对借记卡的《E 法案》，这些法案保护持卡人每次风险赔偿一般不超过 50 美元（Furletti，2005）。中国银联也有类似的保护措施。

5) 商户

本文所指的商户（Merchant）是指接受银行卡支付的商户。他们与收单机构签订协议，安装 POS 终端，并通过收单机构结算服务来收取持卡人的银行划帐资金。上述服务按照不同的行业属性（MCC，Merchant Category Code）向收单机构支付交易手续费（Discount）。商户之所以愿意受理银行卡，早期是由于特定发卡机构能带来更多客户资源，后来银行卡的便捷性成为更主要的因素。随着持卡人群体的日渐扩大，消费者持卡消费习惯的养成，收单服务逐渐成为商户的必备工具。

6) 监管机构

从宏观上来讲，中国所有支付机构及其业务均受央行监管，尤其是中国银联和从事收单业务的非金融支付机构。央行的主要任务是制定行业法规，颁布业务管理办法，维持正常的业务规则和秩序，保护持卡人的合法权益等等。同时中国支付清算协会负责协调产业链各方利益，敦促各方自律。

另外，发卡机构的资质和业务受银监会监管。

7) 支付交易处理支付产业链是一个漫长的链条，大部分支付交易涉及五个参与方（如图 1），少数交易只涉及四个参与方（如图 2）。发卡机构向持卡人发行银行卡、对支付交易给予授权、按期发出账单、收回应收账款。

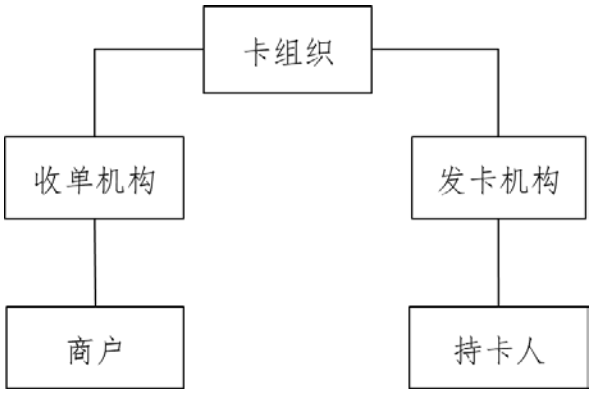

图 1_支付交易处理支付产业链（五方）

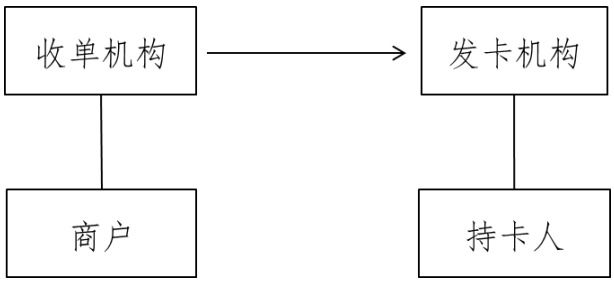

图 2_支付交易处理支付产业链（四方）

交易处理的第一部分称之为联机授权处理(Online Authorization)。联机授权处理包含从店员刷卡至银行同意或拒绝支付请求、返回确认信息的所有过程。收单机构负责签约商户，使商户能够接受银行卡支付。日常运营中，收单机构负责受理 POS 终端发起的支付请求，经前端系统处理，根据读出的卡信息和交易信息，路由至卡组织转接系统或直接转发至发卡机构系统，经发卡机构授权后再按原路返还至商户 POS 终端。

第二部分称之为结算处理（Settlement）。结算处理包括持卡人同意付款后，资金从其发卡机构账户转移至商户账户的过程。交易日终时，卡组织发起批处理，对各成员机构进行账务处理及对账，形成对各成员机构的清算信息，送清算中心轧差(payment netting, closeout netting, cross-product netting)处理；翌日（T+1），收单机构将获得清算资金，之后按总余额（扣除手续费）结算至商户账户。

通常联机授权处理是 IT 过程，考验的是技术和系统能力。结算处理在业务上要复杂许多，因为各方均会有一定比例差错，持卡人和商户也会有众多纠纷，还有大量欺诈行为发生，因此期间包含差错处理、风险后续处理、争议协商等等。

1.3 支付行业对商户欺诈行为预测的要求

银行卡支付交易处理的规则有两个最基本出发点：一是尽量便捷，使得持卡人乐意使用；二是保护持卡人的利益免受损失，使得持卡人无后顾之忧，可以放心使用。不幸的是，前者导致各类欺诈易发；后者意味着风险损失大多必须由商户、收单机构和发卡机构协商承担，而这一切在行业规则中作出了细致规定。

1) 拒付、调单及退单规则

在银行卡交易规则中，为保护持卡人的权益，鼓励持卡人放心使用，规定持卡人具有拒付权力，即当持卡人对购买商品或服务不满意，或对支付交易存疑时，银联规定自交易发生日起的 180 天内（Visa 和 MasterCard 的规则是 90-120 天），可以向发卡机构要求拒绝支付账单。据此，发卡机构通过卡组织向收单机构发起调单，即调取签购单等原始单据，收单机构必须在规定的时间内（银联规则 12-13 天）证明商户的真实性，同时立即向商户收集凭证，以证明交易的真实性。在规定时间内，如果调单举证成功，持卡人须承担账单和调单费用；如果调单失败，发卡机构则有权向商户发起退单，即直接从待结算款项中扣除争议拒付

资金，造成接下来的商户结算短款。如果商户没有履约能力，这损失即由收单机构承担（见《中国银联业务管理办法（二）》）。

以上机制的目的是保护持卡人，但在实际操作中，由于产业链上发卡机构与收单机构市场地位的显著差异，拒付损失或其它欺诈损失很大程度上由发卡机构转向了收单机构，收单机构某种程度上为发卡机构提供了风险交易保险（DeGennaro，2006），背负了巨大的商户欺诈连带责任。

由于自交易日开始 180 天止，收单机构必须承担可能的不法商户的连带风险，因此，收单机构必须具备强大的风险控制能力，尽可能从不同维度来识别出各种交易欺诈，同时也需要对商户资质进行严格甄别，对其行为严密关注。

2) 欺诈的种类

银联给出的欺诈交易定义是指涉嫌使用虚假身份获取银联卡，或冒用他人银联卡（或账户）获取商品或服务的欺骗性交易行为（见《银联管理办法》）。TCS 给出的欺诈定义如下：在持卡人和发卡机构对交易不知情下使用了他人的卡；或者，刷卡人与发卡机构或持卡人没有关联关系，并且没有与持卡人联系或没有还款的意愿（Bhatla，Prabhu，Dua，2003）。

美联储研究人员 Kahn 和 Roberds（2005）认为，欺诈是由于债务人身份虚假造成的债权无法索取的风险，并将欺诈分为三类：一是针对现有卡欺诈，如遗失卡、被盗卡、未达卡等情况；二是新卡欺诈，通过侧录、互联网钓鱼以及二维码等方式获取卡内信息，然后制

作伪卡，甚至直接利用他人身份信息直接申领新卡；三是抵赖，即消费后利用交易规则拒付。

传统收单业务风险大多围绕卡端展开，对每一类欺诈表现出的行为进行分析，通过不断累积经验，建立专家知识库的方式识别欺诈。

3) 商户欺诈

中国银联对商户欺诈的定义是指特约商户在受理银联卡交易时，违规操作、蓄意进行欺诈交易或纵容、包庇、协助银联卡欺诈交易的行为，包括恶意倒闭、套现、洗单、商户虚假申请、侧录、卡号测试、虚假商户、手输卡号等欺诈形式（见《银联管理办法》）。

传统商户一直被视为较少欺诈，理由一是消费者进入实体店后，面对面交流自然会对店家的真实性有直观判断；二是收单机构每年至少一次对实体商户进行巡检，可以判断商户的总体信用情况。因而认为，传统商户只存在一些雇员的不良行为或侧录，构不成规模风险。

然而，随着互联网技术的普及、电子商务的发展、犯罪的团伙化和国际化，情况已发生根本性变化，商户欺诈越来越成为值得关注的重要风险，其表现主要是两个类别：一是非实体商户，通过互联网、移动、二维码、电话、微信等支付方式，通过钓鱼骗取结算资金；二是与持卡人合谋，通过虚假交易（包括虚开价格、现金退货等形式），甚至通过注册虚假商户，从发卡机构获取资金，即所谓套现。

根据西南财经大学甘犁（2014）的报告，中国小微企业（含个体户）的贷款可得性只有 23%，也就意味着小微企业对资金的需求处于严重的饥渴状态。同时中国资金成本存在显著的价格双轨制，民间借贷成本达年利率 30-40%。由于信用卡发卡的普遍及易得性，银行不仅提供最多 56 天的账期，而且还配有灵活的分期还款服务，成本也在年利率 20%左右，因此，套现成为一个普遍的现象，也成为中国收单市场的独特现象。

通常，套现又可以分为三种不同情况：一类是正常经营的商户，他们偶尔套现，以解决商户自己或熟人的短期资金需求，虽不值得称道，但还属良性；二是灰色经营商户，他们利用套现资金从事民间借贷业务，或与高利贷主合谋，这种情况是利用既有规则的金融套利行为；三是犯罪分子经营的恶意套现商户，他们使用假资料与收单机构签定收单合同，一手制作伪卡，一手虚假套现、套积分，最终在骗取大额资金后消失。

通过以上分析，我们不难得出结论，收单机构在欺诈行为的防范上，不仅要对付传统的针对卡端欺诈；在中国更需要对商户行为进行预测，尤其是开拓小微商户收单市场时，需要特别关注以伪卡和套现为主的欺诈行为。

1.4 国际国内的研究概况

应该说银行卡支付领域，尤其是收单的风险研究，长期以来一直是产业层面的课题，较少有学术方面的研究（Rochet&Tirole，2002），原因大概是这个产业发展变化较快，时间

不长，同时也牵涉大量商业机密，因此成果较少发表。尽管如此，本课题所涉领域还是有一定的公开研究成果。

多数银行卡研究报告来自于 Visa、MasterCard、银联等产业公司，以及 AC Nilson 这类市场调查公司。这些报告更多是从产业的角度，给出发展统计数据、行业发展现状、规则研究，从行业高度研究风险和欺诈议题。

美联储是银行卡研究的活跃者，尤其是设立在费城分行的支付卡研究中心。他们更多从行业监管角度研究支付行业，尤其是从法律、规则方面展开对欺诈风险的研究。

欺诈防范和缓释更多是产业界的技术研究对象，在过去几十年里取得了飞速发展，主要集中在以下几类：

1) 简单规则系统（Simple Rule）

简单规则系统主要根据欺诈交易的特征总结而成，是对风控专家的知识和经验模拟，基本方式是依照”if...then...”的准则来过滤交易授权请求。该方法的好处是简单明了，直接在联机交易中进行。缺点在于需要占用额外的交易时间，导致客户等待时间加长，尤其是随着时间的推移，规则会越来越多，计算会越来越复杂；另外，该方法也较多依赖于专家的个人主观判断，频繁导致交易被“误杀”。

2) 风险评分技术（Risk Scoring）

风险评分技术是采用统计模型来识别欺诈交易，根据模型联机计算交易的欺诈似然率。与简单规则 0 和 1 相比，该技术可以给出评分，然后再由人工对最高分的交易做出判读。此技术是很大进步，交易误杀概率大大降低。

3) 神经网络技术（Neural Network）

神经网络技术是评分系统的升级版，由海量历史交易中正常和欺诈交易训练而成，通过模拟人脑的、基于统计知识的模式识别，迅速找到模式与当前交易的相关性。该技术的目标是使系统具有学习能力，试图不断从未来交易中提取规则。

迄今为止，大部分的欺诈技术都是基于交易层面的，而且是嵌入到实时联机交易系统当中，为 IT 系统增加不少负担，增加交易时长。

国内所有银行和银联对风控系统也投入大量人力和物力，目前基本还处于模拟人工专家系统的简单规则系统阶段，鲜有对商户的欺诈行为进行预测。

章节 2　商户欺诈行为模型

2.1 模型的目标和框架

2.1.1 模型目标及方法论

商户欺诈模型的目的就是根据商户的历史数据，对商户未来欺诈的概率进行预测，从而可以提前采取业务处置措施，如关闭交易、冻结结算、降低单笔或日均交易限额、提交黑名单等等。

本课题并不是对某一理论的实证研究(即根据理论提出假设 ,再用数据方法检验假设), 而是运用机器学习的方法，从大数据中直接提取所有可能的特征，并把所有特征设定为自变量，进而利用二元逻辑回归方法寻找这些自变量和观察到欺诈行为之间的关系，建立起对未来欺诈行为的预测模型。论文最终整理了 682 个变量，对于字符型变量，论文在建立模型的初始阶段，会将其简单转化为定性变量；例如：对于省份信息，将省份按照欺诈网点比例从高到低排序；如果网点属于欺诈网点比例较高的前 m 个省份(假定有 n 个省份，则 m 可以取 1,2,...,n-1)，该定性变量值为 1，否则为 0。如果最终某些这样的定性变量被选入了模型，再对这些变量做手工精细调整（例如：将省份分成多类，而不是仅仅两类），观察是否能提高模型效果（模型效果的衡量见第四章 Lorenz 曲线）。模型将给出商户未来欺诈行为的概率，根据概率值给出商户的评级。要求的效果是可以投入到商业应用。

本研究数据样本采集采用面板数据法（Panel Data），即截取一个自然月的所有商户数据作为学习样本，包含更新至该月的静态数据（约 60 万商户）、动态交易数据（约 4000 万笔交易）和外部获取的涉嫌欺诈数据（简称外部数据）。梳理收集到的代理、商户、网点、POS 机终端等静态信息，动态交易信息和历史欺诈行为。根据风险专家和业务专家的经验，从中提炼出可能反映未来欺诈风险的特征变量。所有数据归为算法训练集（60%）和测试集（40%）两部分。建模的过程就是在训练集中通过算法不断学习，寻找最能解释未来欺诈行为的特征变量组合；同时，在测试集上验证这些特征变量解释能力的稳健性，防止过度拟合（over-fitting）的问题。

因变量是根据收集到的外部数据来设定的。假设某商户网点未来一个月发生欺诈行为（即：网点在该月中，有至少一笔交易涉嫌被银联，银行，公安判定为网点需承担全部或部分责任的套现，伪卡盗卡，商户欺诈行为），则因变量值为 1；未来一个月不存在欺诈事件，则因变量为 0。涉及欺诈事件的外部数据来源于三个渠道：一是卡组织风险提示和追偿信息；二是发卡机构的拒付或退单信息；三是公安部门的协查记录。

传统方法中，在整理好特征变量和反映欺诈行为的因变量之后，要从大量数据的统计分析中寻找这些特征变量和因变量之间的关系是非常费时的。得益于 IT 技术的提升和各类统计算法的开发，使得人们能够在具备商业价值的计算时长内，能够发现一个较好的特征变量组合，有效地解释因变量在未来的变化。

本课题对特征变量的选取分粗选和精选两个阶段，以确保选取的特征变量组合从统计意义上具有最强预测能力。

粗选阶段的变量降维有多种方法，例如：支持向量机(Alain Rakotomamonjy，2003)，遗传算法(David J. Fogarty，2012)等等，出于实用性的考虑，本文借鉴苏格兰皇家银行(Moez Hababou, Alec Y. Cheng, and Ray Falk，2006)的做法，文章的方法主要采用单个特征变量的信息值计算和变量的聚类分析；该阶段选择变量的标准是根据聚类的结果，在每类里面选取信息值较大的特征变量，其目的是提取解释维度尽可能多、同时解释能力也较强的起始特征变量集合。

精选阶段尝试 Stepwise 算法和 Lasso(Least Absolute Shrinkage and Selection Regression)算法，通过反复迭代获取最终的特征变量组合；在初始的迭代步骤中，Lasso 算法和 Stepwise 算法的结果基本相同，出于计算效率的考虑，在后续的迭代过程中，则统一采用 Stepwise 算法。精选变量的迭代过程如图 3 所示。

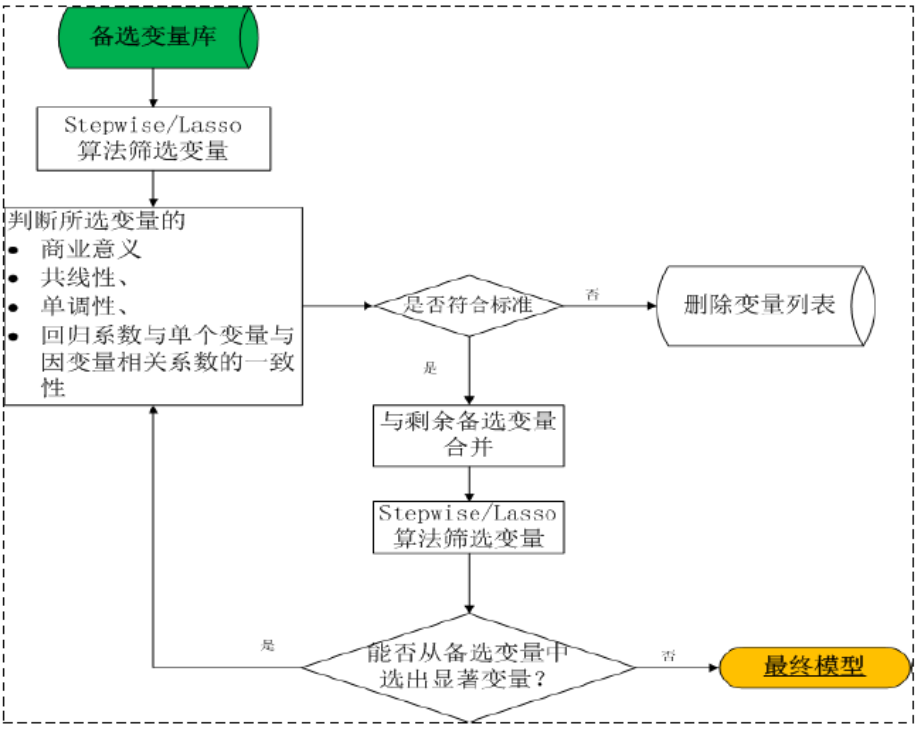

图 3_精选变量的迭代过程

2.1.2 基于收单业务组织的两级模型框架

总体而言，目标模型是为了预测商户的欺诈行为，因此建模之前我们有必要研究商户的组织架构、以及收单业务的拓展模式，以便更准确地，根据商业规律来建立预测模型。在第一章中我们介绍了银行卡产业链以及交易处理的框架，以下我们就收单机构以下产业链条组成作一介绍。

为适应中国的辽阔幅员，针对服务小微商户的商业特点（众多、分散、多位于县、乡等非中心区域），收单机构大多采用了代理模式，即由代理商拓展商户，但仍由收单机构与商

户签订收单合同。商户可能只有一个网点，也可能由多个网点组成，或一个实际控制人同时在多个网点申请安装 POS 终端（见图 4）。

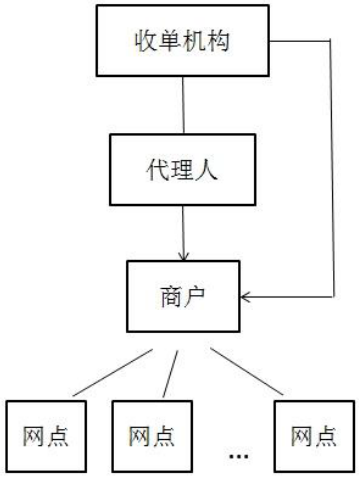

图 4_收单机构代理模式

根据商业特性，商户网点级别可以提供最全面数据（指静态数据、动态交易和外部数据），收单行为也较稳定，因此我们首先建立网点级别的模型；再把网点欺诈概率作为商户模型的自变量，结合商户的静态信息，进而建立商户级别的欺诈模型。

2.1.3 网点分类

根据数据可用性和收单风险专家对于欺诈发生概率的经验，网点可以划分为三个组别：

1) 存量活跃网点，即开通至少一个月以上，且在观察月份有交易；

2) 存量非活跃网点，即开通至少一个月以上，在观察月份没有交易，但之前三个月至少有一笔交易；

3) 新增网点，即签约开通日期在样本观察月份。

对于存量但已经连续 4 个月没有交易的网点，往往在接下来的月份里面也是没有交易的。对于这部分网点，既没有历史的交易信息，也没有未来可观测的数据，本课题不对这些睡眠网点独立建模。

为验证风险专家的判断，我们初步随机抽取了 32.5 万家网点的月截面数据，分析网点在下个月的欺诈比例，发现存量活跃网点、存量非活跃网点和新增网点的欺诈率(在该月中有至少一笔交易涉嫌被银联，银行，公安判定为网点需承担全部或部分责任的套现，伪卡盗卡，商户欺诈行为的网点数和所有网点数的比值)分别为 2.06%、0.19% 和 3.47%。数据验证了风险专家的推测。

据此，我们把总样本按照以上三类网点划分为三类子样本，分三类组别分别进行建模。每类子样本再随机抽取 60%作为训练集，40% 作为测试集。

2.2 网点级别模型介绍

2.2.1 基于信息值和聚类算法的自变量筛选

信息值是用于衡量自变量解释二元因变量能力的一个指标。其计算分为三步：

第一步：针对每一个自变量，先将该自变量按其取值分组(本文将缺失部分为一组，非缺失部分等分为观测值个数相等的 10 组。如果自变量的可能取值小于 10 类，则每类一组)；

第二步：计算每组的证据权重(weight of evidence)，计算公式如下：

$$WOE_i=\ln\left(\frac{\%frauds_i}{\%nonfrauds_i}\right),i=missing,1,2,...,10$$

$\%frauds_i$ 表示在第 i 组中欺诈网点数在所有欺诈网点数中的占比；

同理，$\%nonfrauds_i$ 表示在第 i 组中非欺诈网点数在所有非欺诈网点数中的占比。

第三步，计算该自变量的信息值(information value)，公式如下：

$$IV=\sum_{i=missing,1}^{10}\{(\%frauds_i-\%nonfrauds_i)*WOE_i\}$$

通过对每一变量信息值的计算 ,信息值越大的变量区分二元因变量(即欺诈或非欺诈)的能力越强。

对变量做聚类分析的目的是希望选取尽可能多的解释维度的变量。聚类分析的算法步骤如下：

第一步：对所有自变量做主成分分析(Abdi. H., Williams L.J.，2010)，选择最显著的第一和第二个主成分；

第二步：计算所有变量与第一步选取的两个主成分的相关系数，根据相关性的强弱，将所有变量分为两类。

第三步：将第二步分出来的两类变量分别计算主成分，将所有变量分为四类；如此迭代。

如果满足以下任意一个条件则迭代终止：

1) 该类里面只剩余一个自变量；

2) 与前一轮相比，超过一半的自变量决定系数比(R-Square Ratio)下降。

决定系数比的定义是：

$$\text{R-Square Ratio}(x)=\frac{R_m^2(X)}{Max_{i=1...m-1,m+1,...n}(R_i^2(x))}$$

分子为某自变量 x 由 x 所属的类（假定为第 m 类）的其他自变量做线性回归，得到的决定系数。分母为某自变量 x 由其他类的自变量做线性回归，在这些所有的决定系数中取最大值。

决定系数比越大，说明某自变量能够被该类的变量解释而不能被其他类变量解释的程度越高；也就是从属于该类的程度越大，分类效果越好。

传统的 PCA 方法将所有的样本作为一个整体对待，去寻找一个均方误差最小意义下的最优线性映射投影，但没有考虑类别属性，而它所忽略的投影方向有可能刚好包含了重

要的可分性信息。论文采用逐步 PCA 的方法，即每步只取最显著的两个主成分，方便于逐步观测主成分的商业意义，同时控制迭代停止条件，来决定最终的分类个数；而不是只作一步 PCA，就决定了分类的个数。

最后，基于信息值和聚类算法的自变量筛选的基本步骤如下：

第一步：对所有变量做聚类分析；

第二步：计算所有变量的信息值；

第三步：从每一类中选取信息值较大的一批自变量。

选取的原则有两个：

1) 保证每类中至少有一个变量;

2) 按照每类的信息值之和在所有信息值之和的占比选择相应数量的信息值较大的自变量(即，假定所有变量的信息值之和为 M，某类的所有自变量的信息值之和为 M_i，则在该类里面取信息值大的前 M_i/M 个自变量)

粗选变量阶段的算法结合自变量的商业意义，将从原始的 600 多个变量中选取 100 个左右的变量。

2.2.2 基于 Logistic 回归的建模算法

在完成 2.2.1 的粗选变量后，针对 Logistic 回归(David A. Freedman，2009)，我们运用 Stepwise 算法和 Lasso 算法进一步精选变量。

Stepwise 算法的基本步骤为:

第一步：对每一个自变量做单变量 Logistic 回归，在所有自变量中，选出显著性最高（具有最小 P 值）的那个自变量，记作：X_1；

第二步：选出与 X_1 组合后，显著性第二高的自变量，X_2；如果 X_2 的 P 值大于预先设置的阀值(Significance Level Entry,设定为 0.05)，则停止，模型只有 X_1 一个变量；否者，以 X_1,X_2 为自变量，做 logistic 回归，追溯检验加入 X_2 之后，X_1 还是否显著；

第三步：在剩余变量中，每次选择一个变量，和 X_1,X_2 组合，针对因变量做 logistic 回归。选择最显著的变量，同时追溯检验 X_1,X_2，查看 X_1,X_2 是否还显著，如果其中一个的 P 值大于预先设置的阀值(Significance Level Stay,设定为 0.05)，则删除该变量；

第四步：重复迭代第三步，直到没有剩余变量能够进入组合，在选出的变量中也没有需要删除的则停止。

Stepwise 算法本质上可以看作是贪婪算法；Lasso 算法对其做了修正，在第二步中，并不是寻找与 X_1 最契合的变量；而是寻找与αX_1 最契合的变量，其中 $0<\alpha<=1$。本文以 0.01 为步长，即初始寻找与 $0.01X_1$ 最契合的下一个变量，如果还是 X_1，则调整为 $0.02X_1$，直到找到下一个非 X_1 的显著自变量；显然，Lasso 算法在选择变量的时候更加谨慎，也更有可能找到最优的变量组合，但其计算量也将更大。

在运用提取的自变量建立对因变量的预测模型中，最常用的是线性回归(Linear Regression),其公式如下：

$$Y=\alpha+\sum_{i}\beta_i*X_i+\varepsilon$$

其中 X_i 是第 i 个自变量，而β_i是第 i 个自变量对应的回归系数。回归系数 α，β的确定，普遍使用的是最小二乘法，即寻找最优的 α，β使得残差的平方和$\sum_{k=1}^{n}\varepsilon_k^2$达到最小。

但是，我们对欺诈的预测是二值型的预测，即对网点或商户的欺诈本质上是判定 0 或 1，因此预测函数不是线性的，我们采用 Logistic 回归用于分类（预计分类效果在二维平面上如图 5 所示）。

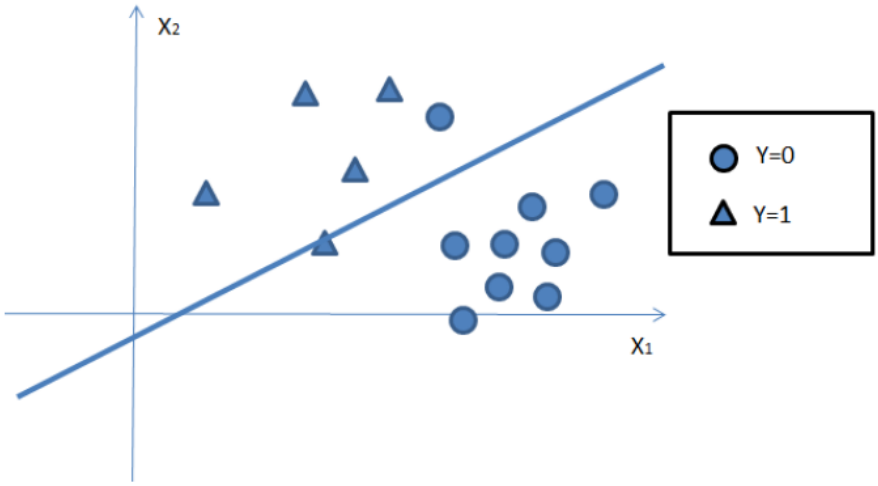

图 5_二维平面上 Logistic 回归的分类效果

Logistic 回归最终是要输出一个概率$(Y=1|X)$，即选定自变量集合 X 后，预测得到的 Y=1 的概率。Logistic 回归的推导需要引入辅助变量$\tilde{Y}=\alpha+\beta X+\varepsilon$，从而将概率计算转化为线性分类问题：

$$P(Y=1|X)=P(\widetilde{Y}>0|X)=P(\alpha+\beta X+\varepsilon>0|X)=P(\varepsilon>-\alpha-\beta X|X)$$

假定ε服从Logit分布，根据其对称性，

$(Y=1|X)=P(\varepsilon>-\alpha-\beta X|X)=P(\varepsilon<\alpha+\beta X|X)=\frac{\exp(1+\alpha+\beta X)}{1+\exp(1+\alpha+\beta X)}$（该函数示意见图 6）。

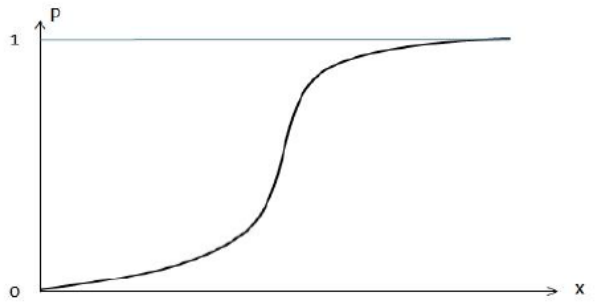

图 6_函数示意图

Logistic 回归系数的确定采用的是极大似然估计 (MaximumLikelihood),即对于每一个观测，由于因变量服从二项分布(1 或者 0)，二项分布的概率就是$\frac{\exp(1+\alpha+\beta X)}{1+\exp(1+\alpha+\beta X)}$，寻找最优的回归系数，保证所有观察值都符合实际发生的情况，也就是极大化

$$\prod_{k=1}^{n}\left(\frac{\exp(1+\alpha+\beta X)}{1+\exp(1+\alpha+\beta X)}\right)^{Y_k}\left(1-\frac{\exp(1+\alpha+\beta X)}{1+\exp(1+\alpha+\beta X)}\right)^{1-Y_k}$$

对上式的每个回归系数求偏导数,置其等于 0，建立非线性联立方程组。求解该方程组采用计算数学的 Newton-Raphson 迭代算法，将由计算机软件 SAS 做自动化计算。

2.2.3 针对三类网点欺诈行为的建模

将近 60 万商户的数据依照三类分组（存量活跃、存量非活跃和新增），编为不同训练数据组进行学习，先后采用上述信息熵提取、聚类算法、Logistic 回归以及相关性分析，得出三个不同组别的网点欺诈模型。

如 2.1.1 节介绍，选择变量和确定模型是个迭代过程，需要比较变量的解释能力和商业含义、变量在模型中的系数符号、变量的显著性水平、变量的单调性、以及模型综合预测能力，需要反复测试后才能完成。为了保证模型的稳定性，还需要进行自变量的共线性检验，保证模型中每个变量的方差膨胀系数小于某个阀值（定为 1.5）。

模型完成后需要输入测试组数据，进行反复验证，确保模型在不同场景下依然足够准确。

2.3 商户级别模型

2.3.1 由网点汇总至商户级别模型的算法

商户可能由多个网点构成，其所属网点的特性组合决定了商户的欺诈特征。因此，在完成网点级别的欺诈预测后，我们可以进一步预测商户级别的欺诈概率。

同样，根据商业逻辑，我们把商户样本分为三个组别：单网点商户、多网点非活跃商户和多网点活跃商户。不言而喻，单网点商户的欺诈概率得分即为该商户的欺诈得分，无需建立新的模型。商户级别模型建立仅针对后两个组别。

商户级别欺诈模型的原理和方法论与网点级别模型类似，也是采用 Logistic 回归方法，不再赘述。

2.3.2 基于商户未来欺诈概率的等级划分

由网点级别汇总至商户级别，根据商户级别模型，我们最终可以计算出商户的欺诈概率。我们将根据概率和商业价值（收益）综合评出五个等级，从而评定出商户优秀、良好、中等、风险、剔出五类商户，以对风险管控和业务拓展提供指导。

章节 3　数据的采集和预处理

3.1 数据来源及可用性分析

数据来源可以分为自有数据和外部数据。

自有数据的部分包括：

1)　代理开户、商户开户、网点开户和 POS 终端申请时备案的静态注册信息。例如，开户、申请的时间、所属地址、网点的注册资本、网点正常营业时间、网点的行业类别、网点身份证显示的所有人年龄、网点的结算类型（对公或对私）等等；对于这部分信息，需要和运营部门沟通，搞清楚哪些信息是申请人、申请公司填写的，哪些是内部操作员填写的；哪些是必填项，哪些是选填项；存储这些信息的系统是否经历过迁移，迁移的时候信息是如何保存的，以保证信息的准确性。

2)　代理、商户和网点在经营过程中，静态信息的变更。例如，代理下属商户数量的变更、商户下属网点的变更、网点所属 POS 机具数量的变更、网点结算账户的变更、网点联系人的变更、网点营业执照的变更、网点行业类别的变更等等。对于这部分数据，需要和运营、技术部门确认信息变更的触发机制，信息变更的权限，是否记录变更的时点，信息变更后历史数据的保存机制。

3)　日常动态交易数据。例如，交易的时间、金额、类别（预授权、消费类）；交易的银联特约商户码（反映交易所属的行业类别）、交易是否发生过退货、交易是否成功、交易

所属网关等等。收集动态交易数据，要特别注意对实时生产系统的影响；一般都是按周或按月，批量下载该周或该月的所有交易数据，而不是实时下载每笔交易数据，以免造成系统崩溃引起操作性风险。因而收集的动态数据是有一定时滞的，目前最快可以做到时滞一天。

外部数据包括：

1) 卡组织（银联）提供的信息，包括卡组织风控系统的预警信息和卡组织向发卡机构确认后反馈的欺诈信息。要获取卡组织风控系统的预警信息，必须成为该卡组织的成员机构，一个基本要求是每年通过该卡组织转接的交易量必须达标并进行相应分润。该类预警信息可以每天在卡组织相关网页下载，时间滞后为 1 天；卡组织向发卡机构确认的欺诈信息，是按季度批量反馈给收单机构的；使用该类数据必须注意时间滞后对数据分析的影响。

2) 发卡机构提供的调单、短款、拒付信息。这类信息由发卡机构不定时地通过文本发送给收单机构，由于每个发卡机构提供的信息和文本格式都不相同。预计信息的采集将是非常费时的。目前的策略是开发规范的整合文本格式，并按周汇总所有发卡机构的调单、短款、拒付文本。另外，按照通行规则，自交易发生日起的 180 天内，持卡人都可以向发卡机构要求拒绝支付账单，因而调单、拒付类的信息最长时滞可以达到半年，这也是使用该类数据的一个潜在问题。

3) 公安协查数据。公安不定时地通过邮件、电话、文本文件等方式要求收单机构参与协查商户，常见的事项有：套现、洗钱、诈骗、赌博、伪卡侧录等等；目前收集这些信息的

方式是按照规范文本版本，按周记录涉案信息，例如：涉案人员、涉案金额、案件时间、地点等等。公安协查有时会涉及多家商户，涉案金额很难清楚归结到每个商户。

3.1.1 商户静态数据信息和动态交易信息采集

在收集到各类静态信息和动态交易信息后，需要对这些信息作加工处理。通过二次开发将这些信息整理成反映风险的特征变量。

例如：单纯从系统中参看网点的所属地区可能无法反映风险内涵，但是如何将网点地区和该网点所属代理地区比对，根据业务逻辑，地区一致的网点由于比较容易受到代理监管，因而风险较低；而地区不一致的网点可能风险较高。

又例如：只关注当月网点的交易量，无法反映风险。但如果将当月交易量和历史月均交易量比较，发现交易量显著增加或显著减少，将体现风险特征；同时将月交易金额和同类（MCC 相同）商户月交易金额比对，如果交易金额显著大于同类交易金额，将体现异常，可能与欺诈风险相关。

分析所采集的数据的可用性和准确性，结合业务专家、风控专家和业界同行的经验。采集原始信息越准确，积累的数据分析经验越充分，二次开发创造的风险特征信息就越有价值越有针对性。

3.1.2 发卡机构、卡组织和公安反馈信息的采集

对于收集的发卡机构、卡组织和公安的原始信息也同样要做二次开发。例如：将卡组织提供的每天风控系统输出汇总到月级别（30 个 EXCEL 汇总至 1 个大的 EXCEL），并统计某月，某网点涉及卡组织疑似欺诈的天数、银行卡数、金额数、交易笔数等等，如果某个网点在某月的欺诈情况较严重，那么在未来其再次欺诈的可能性较高（当然前提是该网点还没有被处置）。由于外部数据获取的方式是多样化的，采集和汇总的工作量是很大的，但这些数据扩大了单个收单机构看待风险的视野，因而将其加入建模是非常必要的。

3.2 数据预处理

3.2.1 基于各数据源的数据整合

在 3.1 节收集各数据源的数据以后，下一步就是将所有的数据源整合为一个风险数据集。

静态数据的基本单元是网点，动态交易数据的基本单元是订单号。连接这两个数据集的逻辑是：根据动态交易数据的订单号找到该笔订单所属的 POS 机逻辑终端号，然后根据逻辑终端号找到所属的网点，最终与静态数据相关联。

外部数据的连接更为复杂，需要根据交易特约商户号，商户简称、交易金额、交易时间、交易卡号等等与内部数据作关联匹配；将这些外部信息与收单机构的某个网点关联起来。

将所有数据源整合在一起以后，可以再做二次开发，整理出一些更有解释力量的变量。例如，将外部信息和内部交易信息组合后，可以统计某月，外部信息反映的欺诈金额在总交易金额中的占比。又例如，将静态信息和动态交易信息组合后，可以统计每笔动态交易的行业类型是否和静态信息中的网点的行业类型匹配。

3.2.2 数据的初步勘探和缺失值处理方法

在梳理完所有的特征变量以后（预计将有近 600 个特征变量），需要对这些数据作初始分析。也就是：判断特征变量的存储类型，是字符型还是数值型。对于字符型变量，需要统计字符型变量可能的取值，每种可能取值的频数分布，缺失率；对于数值型变量，需要给出变量的缺失率，在非缺失的部分计算变量的分布（最小值，最大值，均值，中位数，众数，5 分位数，25 分位数，75 分位数，95 分位数，99 分位数）。对数据的初步勘探，便于我们熟悉构建的数据仓库，同时检查数据创建过程中的纰漏，为后续仿真建模打下坚实的基础。对于缺失值处理，业界有很多种方法；综合考虑精确性和计算复杂度，本论文采用的方法是基于一元回归的算法：

第一步：在非缺失的部分，建立一元线性回归函数关系：

$$X = \alpha + \beta Y$$

第二步：在 X 缺失的部分，由于其对应的 Y 是有值的；直接将 Y 值带入第一步的公式，就是缺失的 X 应该取得值。

章节 4　模型的建立和仿真实验

4.1 因变量的定义

论文的目标是针对各收单网点的欺诈行为建立概率预测模型，并汇总至商户级别，从而得到商户收单业务欺诈概率。各类欺诈行为的具体定义如下：

1)　套现定义：商户与不良持卡人或其他第三方勾结、或商户自身以信用卡为载体，通过虚构交易、虚开价格、现金退货等方式套取现金；

2)　伪卡定义：包括芯片交易方式伪卡欺诈、降级使用交易方式伪卡欺诈和磁条交易方式伪卡欺诈；

3)　盗卡定义：指冒用或盗用持卡人的银联卡进行欺骗交易，包括丢失卡与被盗卡;

4)　非面对面欺诈：欺诈份子窃取卡片主账号、PIN、有效期等账户信息进行冒用，通过 MO/TO、Internet 等非面对面渠道发起的欺诈交易，细分为“互联网欺诈”和“电购/邮购欺诈”两个子类型。

i.　互联网：欺诈分子窃取账户信息后，通过互联网进行欺诈转账或者消费，窃取卡内资金。

ii.　电购/邮购：欺诈分子窃取账户信息后，通过电购或邮购方式进行欺诈交易，窃取卡内资金；

5) 商户欺诈：指特约商户在受理银联卡交易时，违规操作、蓄意进行欺诈交易或纵容、包庇、协助银联卡欺诈交易的行为。包括恶意倒闭、套现、洗单、商户虚假申请、侧录、卡号测试、虚假商户、手输卡号等欺诈形式。

收集这些欺诈行为的来源包括：

1) 银联违规通报报表；

2) 直联行短款/拒付报表；

3) 公安协查日志报表。

根据这些信息可以定义网点级别模型的因变量如下：

对于网点来说，如果它在下一个月发生欺诈行为，则因变量值为 1；未来一个月不存在欺诈事件，则因变量值为 0。该因变量的数据收集渠道包括三类：

1) 银联风控系统通报的在未来一个月有违规行为的网点；

2) 直联银行通报的在未来一个月涉及短款或拒付的网点；

3) 涉及公安协查的网点。

从以上任一渠道通报的网点，因变量记录为 1；否则为 0。

商户级别模型的因变量定义如下：对于商户而言，如果它旗下任何网点在未来 1 个月发生欺诈事件，则该商户因变量值为 1；否则为 0。

4.2 建模样本的筛选和分割

在选取样本时，有必要剔除疑问数据，保证建模数据的纯度。在选取样本时，论文应用了以下筛选条件：

1) 剔除开户日期晚于建模观测月份的网点；

2) 剔除开户日期缺失的网点；

3) 剔除首次交易日期早于开户日期的网点，早于开户日期的交易有可能并不属于当前登记网点；

4) 剔除过往 3 个月和当月均无交易的存量网点，此类网点在接下来的一个月中绝大部分也没有交易，所以涉及本文定义的 “欺诈行为”比例极低，仅 0.02%。由于无法观察其交易行为，因此要将其排除出建模范围。

2014 年 9 月通过筛选规则的网点，将全部作为本次网点级别模型的建模样本；将以上网点汇总至商户级别，所得商户将全部作为商户级别模型的建模样本。2014 年 10 月通过筛选规则的网点和商户，将作为模型的验证样本。

样本分割需要根据商业规律和数据特征来确定，其目的是为了保证分割后组内稳定性(数据可用性，自变量和因变量之间关系的保持) 和组别之间差异性，以避免混合在一起建模造成不必要的相互影响。网点建模样本最终被分割为 3 个组别 (表 1):

序号	组别说明	2014/09		2014/10		备注
		网点数	欺诈比率	网点数	欺诈比率	
1	存量活跃网点：开户日期早于样本观测月，且在观测月有交易；	202,747	2.06%	249,396	1.97%	网点模型 1 (Segment1)
2	存量非活跃网点：开户日期早于样本观测月，在观测月无交易，在观测月之前的 3 个月至少有 1 笔交易；	42,603	0.19%	59,010	0.16%	网点模型 2 (Segment2)
3	新增网点：开户日期是样本观测月。	79,591	3.47%	85,508	3.09%	网点模型 3 (Segment 3)
共计		324,941	2.16%	393,914	1.94%	

表 1_网点建模样本组别分类

各组别根据 6:4 的比例分层抽样为建模数据(Development Sample)和验证数据(In-time Validation Sample)；验证数据用来进行同一时段的数据验证。与此同时，为了检验模型的稳定性，使用同一筛选标准和分组思路的 2014 年 10 月样本，将作为后续时段验证组(Off-time Validation Sample)。图 7 说明了样本的选取过程：

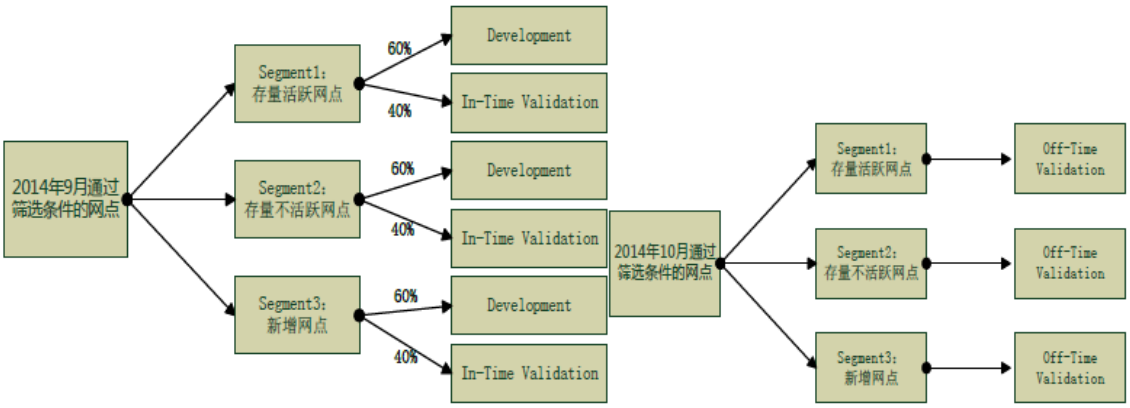

图 7_建模验证组样本的选取过程

商户级别模型是对网点级别模型得分的调整和校正。对于多网点商户来说，旗下各网点的欺诈概率得分不尽相同，这些得分所占的权重也应根据交易金额、网点组别类型进行相应的调整。

将网点级别信息汇总至商户级别时，为了保证商户数据的完整性，筛选条件 2、3、4 剔除掉的网点利用网点级别模型 1 进行预测后，也一并汇入商户级别数据。全部商户样本被分割为 3 个组别（表 2）：

序号	组别说明	2014/09		2014/10		备注
		商户数	欺诈比率	商户数	欺诈比率	
1	单网点商户：旗下仅有 1 个网点。	418,766	1.71%	503,864	1.55%	**无需建模**，商户欺诈概率等于网点欺诈概率
2	多网点非活跃商户：旗下有超过 1 家网点，但全部网点近 4 个月均无交易。	111	0.00%	127	0.00%	若纳入模型样本将掩盖商户的真实行为特征。**不建模**。用商户模型打分后与平均网点得分比较取较大值
3	多网点活跃商户：旗下有超过 1 家网点，至少有 1 家网点近 4 个月有交易。	521	0.96%	706	1.84%	商户模型
共计		419,398	1.71%	504,697	1.55%	

表 2_商户样本组别分类

由于商户模型建模样本仅 521 个，因此不再进行分割。建模样本的选取过程见图 8：

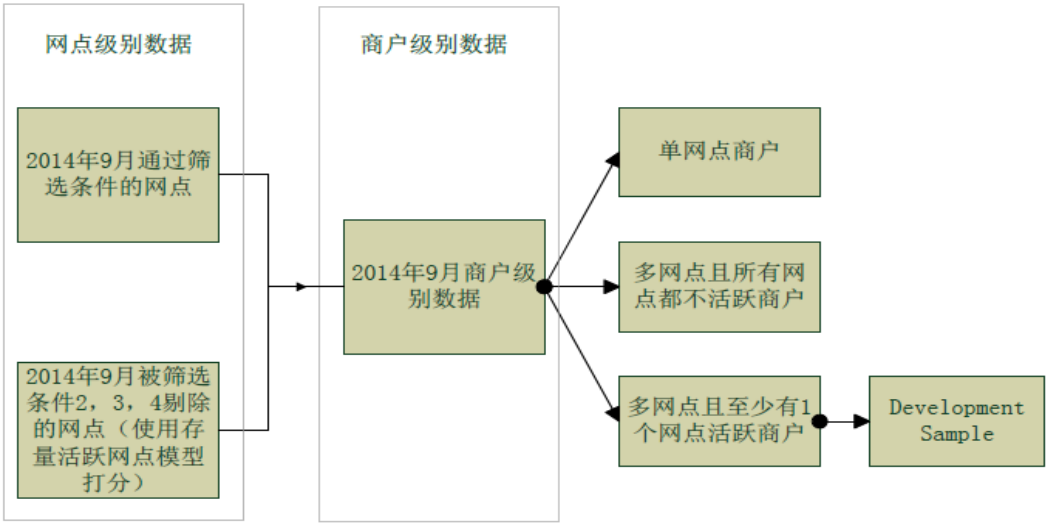

图 8_商户建模样本的选取过程

4.3 自变量的采集和定义

自变量数据的来源在第三章有详细描述，这里用图 9 归纳：

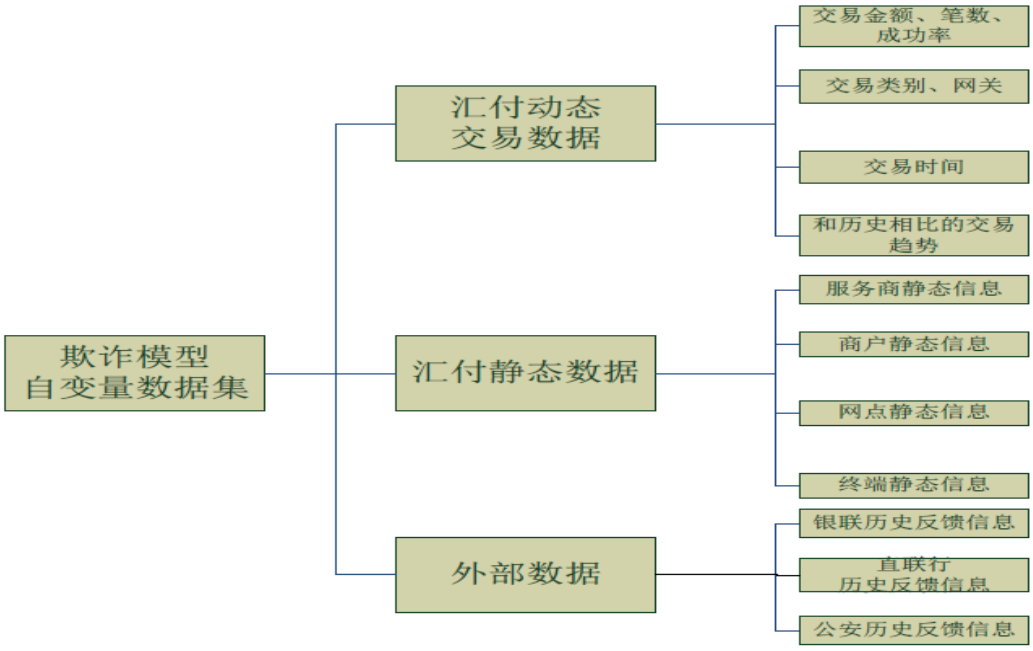

图 9_自变量数据的来源

在采集了这些原始信息后，需要对信息作二次加工，生成有商业意义和预测能力的自变量；做二次加工的角度如下：

地域角度(共 22 个变量)，例如：网点开户省份/地区/邮编与所属代理商/商户省份是否一致；网点开户省份与网点结算账户省份是否一致；所属代理商开户省份与结算账户省份是否一致等等；

时间角度(共 20 个变量)，例如：网点 / 所属服务商开户时间长度；网点法人代表的年龄；网点结算账户最新变更日期距离建模观测月的月份数；网点营业执照有效月份；网点营业时间是否在晚间 9 点以后；网点开户到第一笔交易发生的日期间的间隔，截止到当前月网点未发生交易的月份等等；

运营状态(共 140 个变量)，例如：所属代理商是否关闭，是否缴纳保证金和追加保证金；网点法人证件类型是否为省份证；网点税务登记号是否缺失；网点所属商户结算类型；网点注册资本；网点销售来源；网点结算账户名审核是否有效；网点和所属商户是否使用相同结算银行；网点结算账户是否有被其他网点共用及共用网点数量；网点费率类型；网点当前行业代码和注册行业代码一致性；网点是否经过实名认证；网点管理员在汇付平台留言的次数；网点联系人手机 / 邮箱状态；网点 POS 机具的数量；POS 机具解绑的次数；POS 机具的类型等等；

交易信息(共 277 个变量)，例如：网点历史累计交易金额，历史累计笔均交易金额，历史成功交易金额占比；当月在网点有交易的银行卡数，当月在网点有交易的银行卡数除以三个月的平均交易银行卡数；当月 0 时至 6 时交易金额次数 / 笔数；当月礼拜五和礼拜六的交易金额 / 笔数；当月成功交易金额率除以三个月平均成功交易金额率等等；

历史违规信息(共 223 个变量):例如：网点历史累积退货金额，当月退货金额，当月退货金额除以前三个月(含当月)的平均退货金额；某月交易 MCC 与配置的 MCC 不符的交易次数/金额；银联记录的当月某网点违规天数；银联记录的当月某网点违规卡号的数目；网点历史短款天数，涉及短款的卡数量，短款未追回金额等等。

4.4 模型的训练和展示

整理出近 600 多个可用变量后，一系列变量挑选手段将被应用到变量选择过程中，包括评估变量的 Information Value，变量聚类算法， Stepwise 算法和相关性分析等等(详细算法见第二章)，用来确保从统计意义上进入模型的变量具有最强的预测能力。

选择变量和确定模型是个迭代过程 (见图 3), 需要综合比较变量的解释能力和商业含义，变量在模型中的系数符号，变量的显著性水平和模型的预测表现，反复测试后才能完成。

为了保证模型的稳定性，需要进行自变量的相关性检验并保证变量间的相关性在合理范围内。模型的方差膨胀因子（Variance Inflation Factors，VIF）需要计算并控制在 1.5 以内，方差膨胀因子的计算是将选入模型的某个自变量由其他选入模型的自变量做线性回归计算得到的，计算公式为：

$$VIF(X)=\frac{1}{1-R^2}$$

VIF 等于 1.5 的时候，R^2(即线性回归的判定系数)等于 1/3，也就是某个自变量 1/3 的变动(方差)可以由模型中的其他自变量解释。这个时候论文认为模型自变量之间具有冗余性(多重共线性)，模型自变量需要重新组合。按照模型框架，网点级别模型包含三个组别。

1) 存量活跃网点模型。模型的基本信息见表 3 和表 4：

检验全局零假设:BETA=0			
检验	统计值	自由度	概率
Wald	4246.261	10	<.0001
似然比	5170.227	10	<.0001
评分	10989.6	10	<.0001

表 3_存量活跃网点模型(1)

对模型做全局零假设检验(即能否拒绝所有非常数项的变量其系数同时为 0 的原假设)，论文同时采用了 Wald ,似然比 ,评分(Score Test) 三种方法 ;结果都显示拒绝全局零假设，验证了模型的有效性。

变量名称	变量描述	系数估算	ChiSq概率	标准化系数	VIF	相关性	一致性
Intercept	常数项	-6.922	<.0001	_	_	_	_
CUP_OL_CASE_CNT_L3M	银联记录的前三个月商户违规案例数	0.309	<.0001	0.147	1.09	0.2336	TRUE
gt_18PM_24PM_amt_P80_flag	按MCC分类，如果某商户在当月18：00至24：00的交易金额大于该MCC类别所有商户该时段金额的80分位数则为1，否则为0	1.170	<.0001	0.257	1.15	0.1333	TRUE
SITE_LOW_FEE	网点费率属于低费率(0.38%行业，0.78%,26元封顶行业)	2.047	<.0001	0.558	1.15	0.1057	TRUE
CUP_CHEAT_DAY_CNT_T0	当月银联记录的存在卡诈骗天数	1.195	<.0001	0.081	1.03	0.1730	TRUE
SITE_TAX_CODE_M	网点税务登记号缺失，1；否则，0	0.625	<.0001	0.164	1.12	0.0674	TRUE
site_amt_T0dACC	当月网点交易金额除以历史月均交易金额	0.573	<.0001	0.156	1.06	0.0375	TRUE
AGENT_HRISK_OPEN_PROV	服务商属于高风险省份(贵州，海南，吉林，湖北，广东，江西，重庆)	0.534	<.0001	0.109	1.02	0.0550	TRUE
SITE_MCC_UM_S_TRDCNT_L3M	前三个月交易MCC与配置的MCC不符的成功交易次数	0.002	<.0001	0.047	1.10	0.0377	TRUE
SITE_OPENED_MONTH	网点开户时间长度	-0.028	<.0001	-0.067	1.17	-0.0222	TRUE
short_day_cnt_L3M	三个月（T0，H1，H2）中短款的天数	0.973	0.0006	0.025	1.00	0.006	TRUE

表 4_存量活跃网点模型(2)

自变量的系数是通过极大似然估计方法计算得到的。ChiSq 概率是检验系数是否为 0 的统计量，一般要求小于 0.05(可以理解为拒绝系数等于 0 这个原假设的概率大于 95%)，该模型自变量的 ChiSq 概率显然小于 0.05，显示自变量都是能对因变量产生影响的。标准化系数是将自变量做标准化$\frac{X-MEAN(X)}{STD(X)}$后与因变量做回归，得到的回归系数，将所有自变量都归一化到同一个数量级便于对回归系数做直观比较。

相关性是计算单个自变量与因变量的相关系数，相关系数与回归系数的符号必须保持一致性(即一致性为 TRUE)，确保单个自变量与因变量的关系没有被模型中的其他自变量扭曲。

模型的效果是通过 Lorenz 曲线表现的，如图 10：

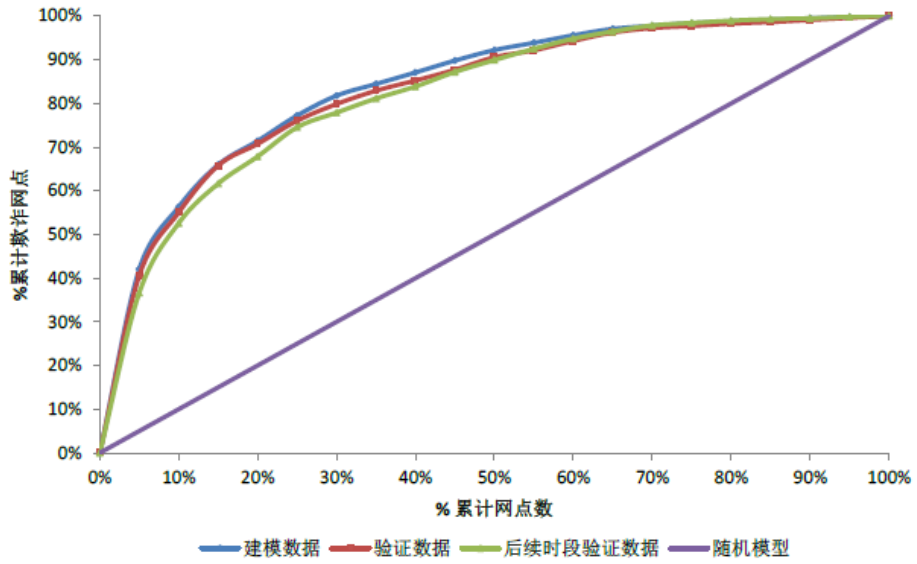

图 10_存量活跃网点模型的 Lorenz 曲线

如果在所有网点中随机抽取 10%的网点(图 10 的横轴)，其结果就是也会随机抽取到10%欺诈网点(纵轴)，所以随机模型在 Lorenz 曲线是一条 45 度的直线；模型可以帮助提高抽取欺诈网点的效率，按模型预测概率从高到低排序，概率最高的 10%网点(图 10 横轴 10%点) 中将会抽取到至少 50%的欺诈网点，显然效率明显提高了。另外模型的稳定性也必须考量，图 10 中可以看出，模型在验证数据和后续时段验证数据中的表现比较一致，因而模型的可用性得到了保证。在这里暂时不检验模型的经济效益；模型的经济效益将在所有子模型都建立完成，将这些模型合并在整体商户上应用时做评估；

2) 存量非活跃网点模型的相关信息见表 5 和表 6：

检验全局零假设:BETA=0			
检验	统计值	自由度	概率
Wald	95.817	6	<.0001
似然比	71.686	6	<.0001
评分	347.503	6	<.0001

表 5_存量非活跃网点模型(1)

对模型做全局零假设检验，Wald，似然比，评分(Score Test)三种方法都拒绝全局零假设，模型有效性得到验证。和存量活跃网点模型一样，模型自变量信息展示如下

变量名称	变量描述	系数估算	ChiSq概率	标准化系数	VIF	相关性	一致性
Intercept	常数项	-6.352	<.0001	.	.	.	.
SHORT_DAY_CNT_1T2	前两个月内短款的天数	2.311	<.0001	0.057	1.00	0.0999	TRUE
SITE_FEE_E_FLAG	费率类型为E-0.78%,26元封顶行业	1.291	<.0001	0.269	1.03	0.0304	TRUE
AGENT_PROV_FLAG	'0035' 福建 '0041'河南 '0042'湖北 '0043'湖南 '0044'广东	1.115	0.0002	0.281	1.01	0.0234	TRUE
AGENT_OPENB_AREA_MATCH	代理商与代理商开户行地区一致,1.	-1.370	0.0003	-0.169	1.02	-0.0301	TRUE
SITE_ORD_AMT_H1_d100k	前一个月网点交易金额除以100k	-1.370	0.0021	0.049	1.04	0.0407	TRUE
SITE_NUM_PNR_MTHLY_H1	前一个月网点ＰＯＳ机具（有交易）数目	0.539	0.0064	0.162	1.03	0.0174	TRUE

表 6_存量非活跃网点模型(2)

可以看到该网点模型的所有指标都符合标准，但选入模型的自变量数量比较少，原因是存量不活跃网点当月没有交易，同时这些网点开户也有一段时间，历史注册信息比较陈旧，中间有变更时收集信息也比较困难。对于该类网点，信息不够充分，建模的难度也提升了。

模型的 Lorenz 曲线，如图 11：

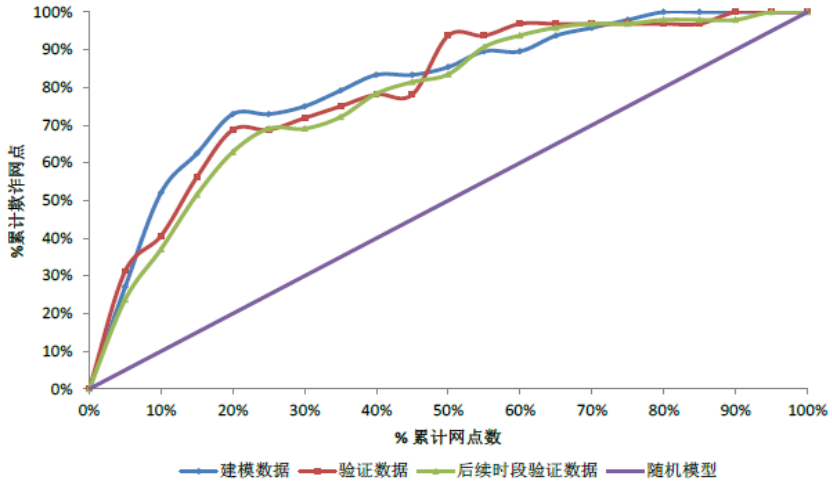

图 11_存量非活跃网点模型的 Lorenz 曲线

由 Lorenz 曲线可以看到，模型的稳定性下降了，原因有两点，第一是可用信息比较少；第二是该类网点的欺诈率(因变量)比较低，加大了模型捕获难度。但模型的表现还是优于随机模型；在前 10%的网点中，可以捕获至少 35%的欺诈网点。

3) 新增网点模型。模型基本信息展示如表 7 和表 8：

检验全局零假设:BETA=0			
检验	统计值	自由度	概率
Wald	2267.459	8	<.0001
似然比	2247.723	8	<.0001
评分	4105.465	8	<.0001

表 7_新增网点模型(1)

对模型做全局零假设检验，Wald，似然比，评分(Score Test)三种方法都拒绝全局零假设，验证了模型的有效性:

变量名称	变量描述	系数估算	ChiSq概率	标准化系数	VIF	相关性	一致性
Intercept	常数项	-6.996	<.0001	_	_	_	_
gt_act_amt_P80_flag	与同MCC类型（交易笔数高于100的MCC)的存量活跃网点比较，若当月网点成功交易金额大于同类型存量活跃网点80百分位，则为1，否则为0	1.977	<.0001	0.2694	1.19678	0.24709	TRUE
AGENT_HRISK_OPEN_PROV	服务商属于高风险省份(吉林，浙江，福建，江西，广东，海南，重庆，贵州）则为1，否则为0	0.813	<.0001	0.2158	1.01577	0.08627	TRUE
SITE_FEE_LOW_FLAG	网点费率属于低费率(0.38%行业，0.78%,26元封顶行业）则为1，否则为0	2.682	<.0001	0.4464	1.02298	0.05918	TRUE
rs_cross_flag_amt_p95	gt_act_amt_P95_flag*SITE_NUM_S_STAT_AMT_T0 其中gt_act_amt_P95_flag是与同MCC类型（交易笔数高于100的MCC）的存量活跃网点比较，若当月网点成功交易金额大于同类型存量活跃网点95百分位，则为1，否则为0； SITE_NUM_S_STAT_AMT_T0是当月网点成功交易金额	0.025	<.0001	0.055	1.14746	0.17479	TRUE
rejected_trans_flag1	若当月失败交易次数超过当月交易次数的20%，则为1，否则为0	0.391	<.0001	0.0979	1.0737	0.087	TRUE
CUP_CHEAT_DAY_CNT_T0	每个月银联记录的存在卡诈骗天数	1.039	<.0001	0.0381	1.05038	0.12418	TRUE
SITE_ACCT_ID_CNT_V2	多个网点（非相同商户）共用一个结算账户，记录网点数	0.092	<.0001	0.0562	1.0104	0.03558	TRUE
rs_avg_risk_amt	高风险笔均交易额：若当月网点成功交易数笔小于等于25笔，且网点笔均成功交易金额大于10000元，则记录笔均成功交易金额；否则为0	0.383	0.0087	0.0217	1.05045	0.04949	TRUE

表 8_新增网点模型(2)

新增网点模型的注册信息比较准确，且一般新开当月都有交易，所以其信息的可用性优于存量非活跃网点；但该类网点缺少自身历史交易数据作为参考，只能通过和同类型的网点历史交易信息作比对，这限制了模型的效力。模型的 Lorenz 曲线，如图 12：

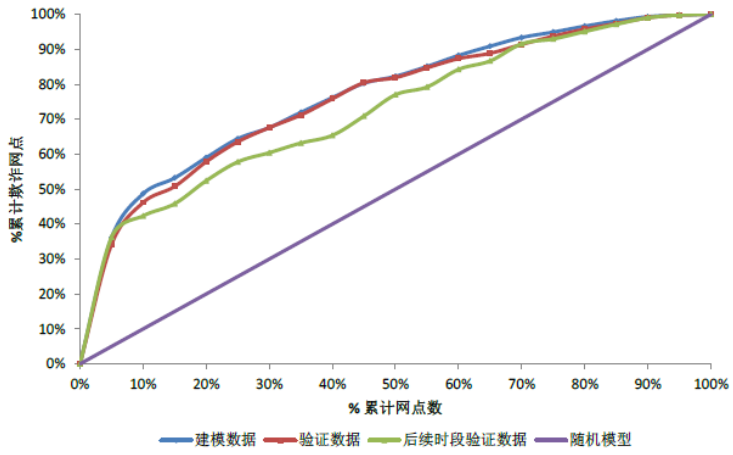

图 12_新增网点模型的 Lorenz 曲线

由 Lorenz 曲线可以看到,模型的表现在后续时段下降。主要是公司处于快速成长期，每月新增的网点在省份，行业，费率等等方面差别都比较大。这导致这个月建立的模型，下个月表现可能就不可靠，论文将该类网点单独建模，也是考虑未来有必要经常更新该类网点模型，而保持存量网点模型不变，提高整体工作效率。模型的效力还是优于随机模型的，在前 10%的网点中(模型标识欺诈概率最高的 10%)中可以至少捕获 40%的欺诈网点。

汇付 99.85%的商户都是单网点商户，直接可以用网点的欺诈概率代表商户的欺诈概率；对于剩下的少量多网点商户，论文出于精确性的考虑，也建立了一个商户级别的模型(见表 9):

变量名称	变量描述	系数估算	ChiSq概率	标准化系数	VIF	相关性	一致性
Intercept	常数项	-44.451	0.9168	_	_	_	_
prob_mer_amtwgt	旗下各网点近3个月交易金额加权平均计算所得bad_rate	10.935	0.0258	4.095	1.00	0.1021	TRUE
mer_site_bal_type	该商户结算方式为“1vs1”则为1，“1vs多”结算为2	14.851	0.0383	2.538	1.10	0.0556	TRUE
inact_2m_10p	若旗下网点超过1.5个月不活跃的比例大于10%则为1，否则为0	10.740	0.0497	0.376	1.10	0.0188	TRUE

表 9_商户级别模型

由于该商户模型的建模数据量很少(521 个观测值)，所以检验统计量的可靠性下降，ChiSq 概率普遍较大。另外，数据集中更新在网点级别，商户级别的数据维护比较少，最终模型选择了 3 个变量。第一个变量是对所有下属网点的欺诈概率以过往三个月的交易金额做加权平均；第二个变量反映如果商户下属网点资金都结算到商户的账户则风险比每个网点自行结算风险高；第三个变量建议如果下属不活跃网点比例高则商户风险高。从商业意义理解都符合常理。

商户级别模型的 Lorenz 曲线，如图 13：

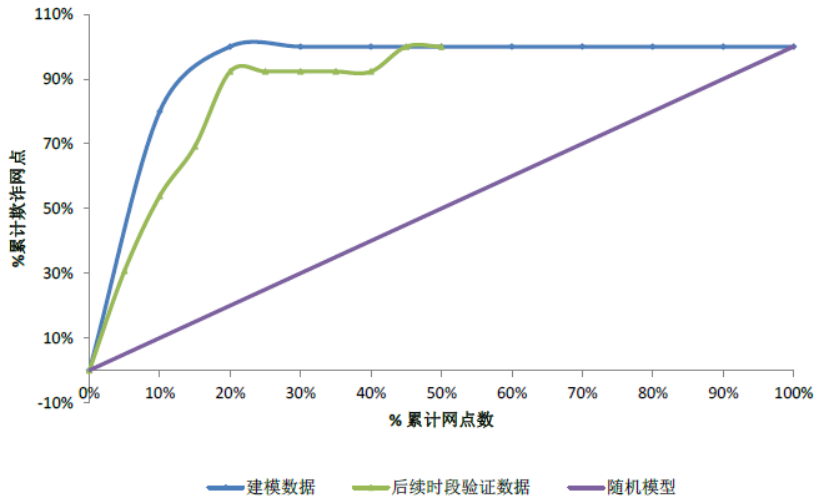

图 13_商户级别模型的 Lorenz 曲线

图 13 中不包含验证数据，这是由于 521 个观测值无法再作分割。后续时段的验证数据显示模型效果有所降低，这个因为建模数据太少，模型在统计意义上不稳定，同时较少数据量的情况下，1 到 2 个欺诈商户就会导致曲线图有所变动。模型的效果还是明显优于随机模型的。

将所有商户合并，查看商户级别上模型的排序能力，见图 14：

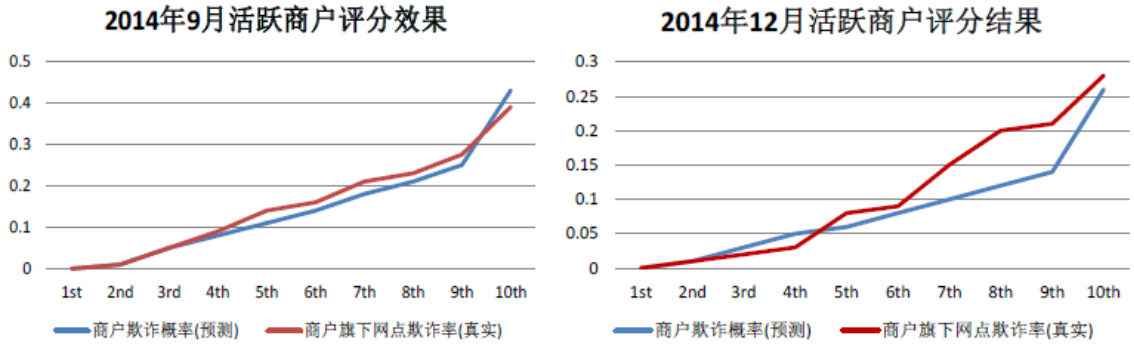

图 14_商户级别上模型的排序能力

可以看到，在九月份的建模数据上，预测和真实值比较接近，预测值（蓝线）显示风险最高的第 10 组，其真实风险（红线）也比较高。在 12 月份的后期检验数据上，点估计的精度下降（预测和真实值差距较大），但是排序能力是保持的，即预期风险较高，其真实风险也较高，在比较延后的时间段上，进一步确认了模型的有效性。

最后我们尝试在 12 月份的数据上分析模型的经济效益，分析经济效益需要基于一些假设，这里包括：

1）如果被模型判断为高风险商户（违规概率比较高），那么就直接关闭该商户； 判断高风险的概率阈值将动态调整。

2）根据经验，假定所有商户给公司带来的利润率为 0.1%。

3）一旦商户发生真实违规行为，允许对这些商户造成的损失做一个平均假设；损失率也将作动态调整。

因此，模型的收益公式可以定为：

模型捕捉欺诈商户交易金额*损失率 - 模型误杀商户交易金额*利润率

在 2015 年 1 月 1 日，按照模型打分，按逾期概率从高到低排序，取 x%的预期高风险商户。然后在 2015 年 2 月份，检查这 x%商户，如果商户在 1 月份中发生了真实欺诈行为，模型有效，这部分商户记为模型捕捉欺诈商户;否则，就是模型误杀商户。具体结果如表 10：

	封杀前10%的高风险商户	封杀前5%的高风险商户	封杀前2%的高风险商户
收益率0.1%，损失率0.2%	¥-2,540,173	¥121,484	¥1,931,742
收益率0.1%，损失率0.3%	¥1,098,395	¥3,089,203	¥4,178,693
收益率0.1%，损失率0.5%	¥8,375,533	¥9,024,641	¥8,672,595
收益率0.1%，损失率1%	¥26,568,378	¥23,863,235	¥19,907,351
收益率0.1%，损失率2%	¥62,954,068	¥53,540,424	¥42,376,862
收益率0.1%，损失率5%	¥172,111,135	¥142,571,991	¥109,785,396

表 10_模型收益效果

可以看到，当公司管理商户能力越强（即商户发生欺诈行为，公司能够尽可能地降低损失率），模型需要关闭的高风险商户数量就越少。在表 10 中如果公司能够将损失率控制在

0.2%，关闭前 10%的商户，则模型甚至给公司带来了负收益，所以模型的使用必须参考公司的具体情况，灵活运用。

4.5 模型的局限和应用

正确应用模型必须满足一定的前提假设条件：

1) 个体的历史表现反映了它未来的行为表现；

2) 样本假设：归入建模组的网点/商户行为表现较为稳定；网点/商户/服务商准入条件、公司风控管理和运营管理基本稳定，与对待建模组样本的标准大致相当；

3) 变量假设：建模中使用的数据和生产系统中的数据基本吻合；模型使用的数据在未来使用中无获取障碍。

模型最直接的应用就是根据商户欺诈概率的高低，将商户切分为多个风险等级。风险等级越高，风险越大的商户群，相对应的风控措施应更为严格，例如加强实地巡检、在风控系统中针对高风险商户设定更严格阀值等从而降低欺诈事件发生的频率。欺诈模型的计算有助于风控人员有针对性地开展工作，提高效率。

章节 5　结论与展望

通过对大样本数据进行采集、清理，进而抽取自变量，我们建立了四类模型，即：存量活跃网点、存量非活跃网点、新增网点级别以及商户级别的欺诈预测模型。通过近一年的数据验证，我们可以得出结论：存量活跃网点和商户级别模型针对欺诈的捕获能力较强，同时稳定可靠；存量非活跃网点和新增网点的捕获能力也较强，但稳定性需要进一步透过未来积累的数据训练加以提高。这些模型将运用于生产系统。

为了建立上述模型，团队开展了历时六个月的研发，除产生上述四个预测模型外，我们还产生了一系列丰富的成果：

- 建立了基于大数据技术的 IT 系统；
- 建立了数据采集、清理和收集的方法和日常机制；
- 初步摸索了一套数据分析和模型建立方法论；
- 初步搭建了覆盖上述工作的团队。

基于以上结论和成果，笔者对后续研究或管理细化作以下展望：

5.1 持续改进模型

5.1.1 模型框架的调整

出于开发时间的限制，目前模型并未区分商户欺诈的类别，比较精确的做法是针对每一类别的欺诈单独开发相应的模型，例如：套现行为有一套模型，伪卡盗卡有一套模型，商户欺诈有一套模型。这样的模型将更有针对性，使用上也更灵活。

另外，模型对商户欺诈行为的损失程度没有衡量，即只要商户欺诈就认定模型的目标商户。目前模型这样做的原因是为了避免后续的人为干扰和不确定因素，例如：损失金额小有可能是因为公司的催收人员的努力而不是商户本身质量好；损失的金额可能涉及到多个商户，难以区分每个商户造成的损失金额等等。如果以后数据收集条件进一步完善，作者将采用加权 Logistic 回归的算法，即建模时每个商户附加损失金额作为权重考虑到模型的因变量中去，提高模型的经济价值。

5.1.2 增加宏观经济相关自变量

根据商业经验，商户欺诈的主要目的是对信用卡的恶意不还款，以及以套现为目的的虚假交易，而后者越来越成为欺诈的主要形式，其深层原因是小微企业、个体户和消费者难于从正规银行获取贷款，而通过信用卡套现方式是变相的、从银行获取的套利，是中国现有商业环境下“特殊的、自我实现的贷款方式”。

因此，和很多贷款类似，套现受经济周期、行业景气、地域经济等事件影响。本研究自变量中只有地域相关信息，如何从其它数据中将其它重要信息加入模型中，丰富自变量内容，是下一步值得研究课题。

5.1.3 时间序列相关事件影响因素

本研究是按日历月份为面板截取数据，存在以下问题：一是业务开展时间并不长，故选取验证的样本时间长度有限，需要持续的跟进，作进一步的模型调试和验证；二是收单业务处于快速的变化之中，技术革命、监管政策、定价机制、行业竞争、对外资开放等对行业格局有重大影响的因素日新月异，模型可能需要快速适用新的形势；三是犯罪分子或套利者利用新技术、新规则不断地发明新的欺诈模式，而且由于互联网的传播速度，欺诈的破坏力不断升级，因此模型的变化速度要非常迅速，适用收单欺诈创新的节奏。

5.2 完善管理办法

再好的模型，如果没有管理措施配套，也将沦为“花瓶”。欺诈模型的应用是一项复杂的管理体系，主要原因是：收单机构要成为一个健康的、可持续发展的公司，必须让商户、持卡人、银联、发卡机构同时满意，也必须兼顾满足监管机构合规的要求，树立具有社会责任的企业形象。

如果运用模型简单粗暴管理商户，那么一定会引起商户和持卡人的客户体验下降，引起客户的反弹；如果执行得过松，那么银联和发卡机构一定会通过“追偿”方式把损失转嫁到收单方；以上两个极端都最终会导致监管机构的介入，可能招致合规处罚。

因此，设计好一套完整的管理方法和流程是模型应用的关键，其重要性不亚于模型研究，具体包括以下七个方面：

1. 商户风险评分

2. 准入条件及审批规则

3. 风险通报机制

4. 商户巡检机制

5. 处置机制

6. 黑名单共享制度（含商户、代理、业务员）

7. 代理评级制度

5.3 进一步的研究展望

基于本课题已经建立起的数据基础，因此完全可以开展进一步的数据挖掘工作，对商户的行为作出分析判断，并据此挖掘商户的潜在商业价值。同时，由于商户的规模在不断扩大，收单商户的行为某种程度已经反映了中国中小微商户或企业的经营状况，也能反映他

们对于金融服务可得性的需求程度，因此数据经进一步加工，可以编制有价值的数据分析报告，也可以形成对相关监管机构或政府部门的政策建议依据。

以下是进一步有待开展的课题。

5.3.1 小微商户的信用评分

小微商户是金融服务的难题，既是一片没有被传统金融机构服务的蓝海市场，又是充满风险、难于驾驭的领地，究其原因，还是因为针对小微商户的信息缺乏，信息收集加工成本相对较高。因此利用收单业务积累的数据，分析小微商户的经营状况，摸索小微商户业主的诚信记录，开展信用评分研究，不失为一条路径。

5.3.2 交叉营销研究

交叉营销是金融服务的重要推广方式，通过某项基础的高频金融服务，金融机构获取广大客户资源，建立与客户的紧密联系，从而在分析客户行为的基础上，提供其它金融服务。

同时，大多数小微商户长期缺乏金融服务，之前也没有与金融机构建立稳定持续的业务关系。借助收单业务，小微商户大量信息开始由收单机构逐步收集整理。通过商户支付业务带来的交易数据和静态信息,借助大数据技术，可以基本判断商户的行为（包含商品服务品类、经营周期、流水、经济收入及毛利状况），因此推断流动性需求、风险承受能力、信

用状况等指标，从而推广或定制金融类产品，既发现金融利基市场，也提升支付业务的粘性。

围绕着小微商户的特征和业主的需求可以设计丰富的金融产品,具体包括以下可能性:

1. 用于周转的小额贷款

2. 理财产品

3. 消费信贷

4. 预付卡产品

5. 会员卡以及积分系统

5.3.3 小微商户相关数据研究报告

小微商户的支付信息反映了大量经营信息，也间接地透露了许多行业和地域的发展状况，经分析整理，可以形成不同地域和行业的数据报告。我们结合西南财经大学中国家庭金融调查中心的数据，联合编制《中国小微企业发展指数报告》，按季发布，已成为众多金融机构和政府部门研究小微企业、就业状况、创业环境、地域经济等课题的重要数据来源。

参考资料

Peter Harrington. 2013. *Machine Learning in Action*.

Jean-Charles Rochet. Jean Tirole. 2002. *Cooperation among Competitors: Some Economics of Payment Card Associations*, RAND Journal of Economics, 549-570.

Tej Paul Bhatla. Vikram Prabhu. Amit Dua. 2003. *Understanding Credit Card Frauds*, Card Business Review.

Michael Cornish. Kathleen Delpha. Mary Erslon. *MasterCard International Security and Risk Management*.

Mark Furletti. 2005. *The Laws, Regulations, and Industry Practices That Protect Consumers Who Use Electronic Payment Systems: Policy Considerations, Payment Cards Center Discussion Paper*, Federal Reserve Bank of Philadelphia.

Ramond P. Degennaro. 2006. *Merchant Acquirers and Payment Card Processors: A Look inside the Black Box*, Economic Review, 27-42.

Moez Hababou. Alec Y. Cheng. Ray Falk. 2006. *Variable Selection in the Credit Card Industry*, Northeast SAS Users Group.

Alain Rakotomamonjy.2003.*Variable Selection Using SVM-based Criteria*, Journal of Machine Learning Research, 1357-1370.

David J. Fogarty. 2012. Using Genetic Algorithms for Credit Scoring System Maintenance Functions, *International Journal of Artificial Intelligence & Applications*, Vol.3, No.6.

Abdi. H. and WilliamsL.J.2010.Principal Component Analysis.*Wiley Interdisciplinary Reviews: Computational Statistics*, 2: 433–459.

David A. Freedman. 2009. *Statistical Models: Theory and Practice*. Cambridge University Press. 128.

甘犁，2014，*小微企业金融指数中国银联业务管理办法（二）*