— 中 国 数 据 分 析 行 业 核 心 刊 物 —

# 中国数据分析

CHINA DATA ANALYSIS 数据分析·因你而不凡

关注CPDA数据说

《中国数据分析》行业特刊
2021年第01期 总第45期（季刊）
咨询热线：400-050-6600
http://www.chinacpda.org/
投稿请至 xiehui@chinacpda.org

中国商业联合会数据分析专业委员会 主办

# 我们的征途是星辰大海
## ——2021 再出发

回首来时路，2020年是极不平凡的一年，更是刻骨铭心的一年。“青松寒不落，碧海阔逾澄。”历经艰难险阻，付出巨大努力，诸多奋斗目标在2020年画上句号。我们不仅见证历史，更在创造历史。艰难方显勇毅，磨砺始得玉成。

在新冠肺炎疫情的巨大冲击之下，各类衍生风险不容忽视，各种“黑天鹅”“灰犀牛”层出不穷，我们在充满不确定性的环境中谋求发展。前进道路上，不可能一马平川，我们必须付出更为艰苦的努力，唯有奋斗才能成就。在危机中育先机，于变局中开新局。新征程上，我们更需要增强危中有机、化危为机的机遇意识。2021年到2035年，是国家规划新的重要阶段，新发展格局、新技术、新产品、新业态、新模式……而创新发展是构建新发展格局的必然选择。

年初，在协会的发展战略座谈会上，大家一起畅聊过往、展望未来……

2020年，我们响应政府号召，全面配合疫情防控工作，克服困难，在危机中寻找生机，在人才培养领域改革、创新的实施,极大满足了数据分析人才的需求，同时积极发挥自身在大数据方面的资源、平台、技术和人才等优势，疫情期间迅速启动“CPDA智慧教室”，以信息化手段助力网络教学工作，并且将传统的沙龙活动、培训转至线上，通过直播等各种形式的展现，有效帮助数据分析师不断提升能力，助力挖掘大数据潜在规律及价值，在数据分析行业中发挥重要作用。

在国家大力主导大数据产业发展及大数据人才培养的背景下，作为全国数据分析行业管理协会，已培养数万名数据分析师和近百家数据分析师事务所，他们以其专业性践行数据强国发展战略，帮助广大企、事业单位实现数字化转型新格局。为了更好的促进国家大数据人才培养战略的实施，近年来我会深入评估高校相关专业，对符合数据分析行业人才标准体系的高校专业开放考试条件，加快培养社会急需的具备大数据处理及分析能力的高级复合型人才。而对于明天，在培养大数据人才的这条道路上，我们依然不忘初心，深耕细作，不断完善人才发展，不断推进素质教育，创新教育方法，提高人才培养质量，努力形成有利于创新人才成长的育人环境。

同时，今年我们组织开展的行业标准化制定工作，通知一经发布，立刻受到社会各大数据公司及数据分析行业从业人员的高度重视与积极响应。规范行业从业单位合规经营、规范其行为、倡导行业自律、维护行业正当竞争，营造一个公正、有序的发展平台与空间……这是一项对社会和行业都具有重大意义的事情。

2021年， 农历辛丑牛年，是辛勤耕耘的一年，“牛”象征着拼搏实干的精神。胸中若有凌云志，不带扬鞭自奋蹄。我们也将引领行业，不断创新、前行，虽然前进道路不会一帆风顺，但惟其艰难，才更显勇毅；惟其笃行，才弥足珍贵。让我们永葆初心、牢记使命，乘风破浪、扬帆远航。

我们的征途是星辰大海，2021，再出发！

中国商业联合会数据分析专业委员会

# 本期目录 CONTENTS

卷首语
01 我们的征途是星辰大海——2021 再出发

协会动态
03 恭喜！以下企业和个人入选参加中国数据分析行业标准化制定工作
04 探讨你该如何从数据分析中体现工作价值
05 用数据思维做好用户画像为企业赋能

政策向导
06 贵州：大力推动大数据与实体经济融合发展
07 公安大数据建设再次被强调
08 人民日报：套路多损害消费者利益维权难大数据“杀熟”必须重罚
10 厦门：信用大数据，让守信者处处便利
12 税收大数据显示：2020年企业销售收入同比增长6%

行业动态
13 金融科技企业50报告：AI、大数据等底层技术企业大幅增加
15 2021年大数据分析的5大挑战
17 数据产品经理必看！数据看板搭建攻略
22 战"疫"启示录：追本溯源，求解智慧城市建设最大公约数

学"数"交流
25 企业如何做生产设备更新决策
26 描述统计分析应用的感触
30 基于大数据的用户流失及预警模型的构建研究
33 实战案例|泰坦尼克号乘客生存预测
36 64个数据分析常用语

事务所专栏
40 中国数据分析行业会员单位——数据分析师事务所介绍

主办单位
中国商业联合会数据分析专业委员会

编委成员
会员处 李苗苗

出版时间
2021年03月出版 <总第45期>

美工设计
市场处 崔峻珩

联系我们
中国商业联合会数据分析专业委员会
地址: 北京市朝阳区朝外SOHO-C座9层
电话: 400-050-6600 / 010-59000991转652
传真: 010-59000991转 607

欢迎广大读者踊跃投稿，内容包括学术观点、教学体验、教学活动、学习感悟、实战经验、随笔文章等。
稿件附图格式为JPG或TIFF格式，大于1M，分辨率在300dpi以上。
感谢您对《中国数据分析》的支持！ 投稿邮箱：xiehui@chinacpda.org

# / 恭喜！以下企业和个人入选参加中国数据分析行业标准化制定工作 /

来源 / 中国商业联合会数据分析专业委员会　编辑 / 协会会员处　李苗苗　日期 / 2021-01

中国商业联合会数据分析专业委员会

中商联数据分析委〔2021〕01号

入选参加中国数据分析行业标准化制定工作

名单公布

我会组织的行业标准化制定工作，目的在于规范从事数据分析相关业务的大数据公司合规经营，规范其服务行为、倡导行业自律、维护行业正当竞争，给从业大数据公司营造一个公证、有序的发展平台与空间。

我会自2020年12月发布《数据分析行业业务范围及收费标准参考意见》邀约通知以后，受到社会广大大数据公司的高度重视与积极响应，此意见将作为数据分析行业从业参考依据，将通过相关新闻媒体对社会进行发布，也将在中国数据分析行业峰会、行业刊物及各种资源平台进行传播，具有广泛的社会影响力和行业影响力。现将入选名单公布如下：

一、入选的单位有（排名不分先后）：

湖南翰林数据分析师事务所有限公司、北京鼎盛恒信项目数据分析师事务所有限公司、海南智企数据分析师事务所有限公司、山东智谷数据分析师事务所有限公司、中易数安（北京）项目数据分析师事务所有限公司、北京中辉智诚数据分析师事务所有限公司、帆软软件有限公司、北京永洪商智科技有限公司、上海荣增信息科技有限公司、贵州佰仕佳科技集团股份有限公司、北京小阳科技有限公司、北京家校帮科技有限公司、山东邦维信息科技有限公司、日照山川电子信息技术有限公司、西安九天数智信息科技有限公司、北京派可数据科技有限公司。

二、入选的个人有（排名不分先后）：

胡旭、李妹、高松、吕雪芬、胡明、范煜、许建真、陈虹坚。

恭喜以上单位及个人！

中国商业联合会数据分析专业委员会

2021年1月26日

中国商业联合会数据分析专业委员会于2020年12月15日发布了《关于邀请参加数据分析行业标准化工作制定的通知》，通知一经发布，立刻受到社会各大数据公司及数据分析行业从业人员的高度重视与积极响应。

短短两周协会收到了百余封申请表踊跃参加评选。经行业协会对参评人员的行业规范性、知名度、服务体系、行业经验等层层资质审核，现已公布入选名单，下一阶段行业协会将根据标准化研究方向，组织参评人员共同讨论确定具体内容。

**企业入选名单：**

- 湖南翰林数据分析师事务所有限公司
- 北京鼎盛恒信项目数据分析师事务所有限公司
- 海南智企数据分析师事务所有限公司
- 山东智谷数据分析师事务所有限公司
- 中易数安(北京)项目数据分析师事务所有限公司
- 北京中辉智诚数据分析师事务所有限公司
- 帆软软件有限公司
- 北京永洪商智科技有限公司
- 上海荣增信息科技有限公司
- 贵州佰仕佳科技集团股份有限公司
- 北京小阳科技有限公司
- 北京家校帮科技有限公司
- 山东邦维信息科技有限公司
- 日照山川电子信息技术有限公司
- 西安九天数智信息科技有限公司
- 北京派可数据科技有限公司

（以上排名不分先后）

**个人入选名单：**

胡旭、李妹、高松、吕雪芬、胡明、范煜、许建真、陈虹坚（排名不分先后）

# / 探讨你该如何从数据分析中体现工作价值 /

来源 / 中国商业联合会数据分析专业委员会　编辑 / 协会会员处　李苗苗　日期 / 2021-01

在工作中总是遇到这样的问题：工作中遇到大量的数据信息却无从下手？面对多种复杂的数据工作时总是出现问题？针对多种客户群体不知道怎样去分析客户的需求？在企业发展数据调研中，不知道怎么抓取精准数据？

2021年1月19日，由中国商业联合会数据分析专业委员会科研中心课程设计师、特聘讲师蒲慧丽老师、张磊老师做客直播间，和大家探讨“你该如何从数据分析中体现工作价值”。

直播期间，老师向大家展现数据分析师的实际工作流程、未来成长途径是怎样的。并且利用阿里巴巴和各大企业、机构数据库的数据，和大家探讨其中的价值。数据分析在实际工作问题中得到应用，又有哪些分析思路可以解决数据问题。同时，老师还和大家远程连线，一对一地深入探讨在实际分析数据方面的新思路。

**讲师介绍**

薄慧丽

中国商业联合会数据分析专业委员会科研中心 特聘讲师

太原理工大学统计学硕士，曾任职今日头条商业化北京分公司、蛋壳公寓，负责业务团队的数据监控与决策支持以及成本模型的离线开发优化。熟练使用数据分析工具Excel和R语言，具有丰富的数据分析工作实战操作经验与丰富的教学经验。

张磊

中国商业联合会数据分析专业委员会科研中心 特聘讲师

研究方向为大数据精准营销分析与建模，潜客场景研究，教育地产医疗等等行业数据分析研究。教学风格深入浅出，对Python数据分析、机器学习等相关的内容具有多年的实践经验。疫情期间，数据分析的应用价值不断凸显，CPDA数据分析师们为社会、企业的数字化进程贡献着自己的力量。

# / 用数据思维做好用户画像为企业赋能 /

来源 / 中国商业联合会数据分析专业委员会　编辑 / 协会会员处　李苗苗　日期 / 2021-01

现今各行各业都对数据分析从业者有大量的需求，疫情期间全国各地都在运用大数据技术统计监测疫情防控工作。数据分析、用户画像，也在疫情追踪溯源、路径传播、发展模型预测、资源调配等领域得到广泛应用。

猎聘2019年大数据人才就业趋势报告显示：中国大数据人才缺口高达150万。台湾《旺报》发布的《2018全球大数据发展分析报告》指出，大陆针对“互联网+”和“大数据+”的融合创新，积累了丰富的数据资源，相关人才占据全球半数之多，从世界范围来看中国大陆大数据相关人才数量虽遥遥领先于各国，但仍存在较大人才缺口。

面对最近经常有各行各业人士咨询与增长业务相关的问题，比如如何实现项目增长等，我们都会先反问他们一个问题——你们的用户画像是怎样的？

久而久之我们发现，很多公司其实都没有一个明确的用户画像，也就是说，他们自己都不清楚自己的用户是谁。更有甚者，连什么是用户画像都不清楚。

**什么是用户画像？如何做好用户画像？**

2021年1月28日，由中国商业联合会数据分析专业委员会科研中心课程设计师、特聘讲师蒲慧丽老师、张磊老师做客直播间，和大家探讨“如何用数据思维做好用户画像为企业赋能”。

直播期间老师带我们深入探讨了用户画像的概念和作用，以及为什么企业需要用户画像，我们如何运用用户画像快速构建完善的客户增长体系等问题。

并教大家利用数据分析思维构建用户画像，探讨高效建立用户画像标签体系，以及运用RFM模型衡量客户价值和客户创利能力。

# / 贵州：大力推动大数据与实体经济融合发展 /

来源 / 贵州日报　编辑 / 协会会员处　李苗苗　日期 / 2021-02

新时代呼唤新经济，互联网时代，产业数字化正成为经济高质量发展的重要途径。今年的政府工作报告中专门提出，要围绕建设国家大数据(贵州)综合试验区，狠抓大数据与实体经济融合发展。

近年来，贵州深入实施“万企融合”大行动，推动大数据和十大工业产业、服务业创新发展十大工程深度融合，有力促进全省经济发展质量变革、效率变革、动力变革。数据显示，贵州实施“万企融合”大行动以来，已打造形成了305个标杆项目、3262个示范项目，带动5946户实体经济企业与大数据深度融合。大数据与实体经济融合发展，正成为引领贵州经济结构转型升级的重要推动力。

**工业：智能化改造让企业“脱胎换骨”**

2月1日，记者在位于贵阳经开区的贵州贵材创新科技股份有限公司看到，一根根长方形的铝型材板通过预处理、脱脂、喷绘、固化后，由立式粉末喷涂滑出台缓缓运送到辅助下料的工人面前，工人将成型的铝型材板分拣到操作台上，在传输机床的协助下，1分钟就有35支成型铝材下线，产值可达1900元。

“以前人工操作，一分钟能完成10根铝型材喷绘就算高效了。”该公司副总经理李衍雷说，得益于自动化，公司生产效率大幅提升，年产能从1万吨增加到5万吨，产品合格率从以前的82%提升到95%以上。

无独有偶，贵州雅光电子科技股份有限公司也是生产线智能化的受益者。

该公司董事长席建军向记者介绍，2017年，雅光电子按照以主业为轴，往高端化、高技术、高品牌、高品质方向转型的原则，投资1500万元建设了一条世界一流水平的高端主机汽车二极管智能化生产线，建立了OE(自动化示范生产线)、PDM(研发数据管理平台)、ERP(企业资源计划系统改造)、CI(数据采集应用平台)与BI(数据分析平台)等智能制造系统。一条生产线上的员工由38人减至12人，产品产量提升58%，优品率提高33%，经济效益提高了15%。

大数据与工业的融合背后是更高的产能、更有效的劳动。数据显示，2020年全省大数据与实体经济深度融合指数达到41.1，比2019年提升1.6，整体融合进程已初步进入中级阶段。

**农业：产销智慧对接推动“黔货出山”**

“这是全省首个整合蔬菜农产品供应端和销售端信息的B2B平台——黔菜网·贵州农产品大数据平台。”

指着屏幕上不断滚动更新的数字，贵州蔬菜集团董事长张文贵对记者说，平台现有基地动态、扶贫专区、供应大厅、采购大厅、冷链配送、包装包材、黔城超市、价格行情共8大

功能，全省各地的蔬菜种了哪些品种?什么时候上市?在平台上都能一目了然。不仅如此，还可以通过分析数据反过来指导生产，从而加强内部产业结构调整，把产销对接这个纽带真正建立起来。

截至目前，黔菜网·贵州农产品大数据平台已覆盖全省88个县(区、市)、3731个基地、315.62万亩，注册用户5994人，线上店铺971个，商品总数230238件，总成交金额达66796万元。

黔菜网·贵州农产品大数据平台只是贵州不断加强农业与大数据融合的案例之一。

近年来，贵州充分利用大数据发展先行优势，在深化农业产销智慧对接、推动农业生产智能化等方面不断取得重大突破，大数据+农业融合覆盖面不断扩大，各类智慧农业成绩斐然。

“一码贵州”智慧商务大数据平台自2020年7月上线以来，累计入驻企业3.33万家，上线产品7.23万个，完成订单65.77万笔，实现贵州产品销售40.14亿元;“辣椒云”“茶云”“猪联网”“农经云”等一批行业特色平台累计交易额超过70亿元;农产品销售智能终端进入社区，建成农产品仓储保鲜冷链设施项目327个。

**服务业：新业态创造“智慧生活”**

“听说今天有专家过来给我们看病，而且来了就不走喽。”“真的假的，我们赶紧过去看看”……在罗甸县茂井镇布江村，听说村里来了大医院的专家，村民纷纷赶往村卫生室，都想让专家给自己瞧一瞧。

来到村卫生室，找了一圈也没看到专家的身影，只多了台电脑显示器。面对满脸疑惑的村民，布江村驻村医生王家语赶紧向大家解释，“大家别急，专家就在显示器里。”

原来，为改善贫困地区老百姓就医状况，贵阳朗玛信息技术股份有限公司在布江村建立了远程互联网医疗点，让当地村民能够“足不出村”享受到省级优质医疗资源，解决贫困地区“看病难、看病贵”的问题。

突如其来的新冠肺炎疫情，为贵州数字经济带来了一个意想不到的契机。全省在线医疗、线上办公、网上购物、线上旅游、平台经济等新业态新模式频频涌现。

去年5月19日发布的“一码游贵州”全域智慧旅游平台就是其中之一。这个涵盖千种景观、千种风味、千种风物，为广大游客提供“吃、住、行、游、购、娱”等方面智慧旅游服务的平台，一经推出就深受文旅企业和游客的欢迎。

截至目前，平台已入驻景区492家、酒店21985家、商户23414家，累积上架特色商品24000多种，总访问量近1亿人次，累计使用用户近1000万人。

# / 公安大数据建设再次被强调 /

来源 / 东方证券　编辑 / 协会会员处　李苗苗　日期 / 2021-02

1月18日至1月19日，国务委员、公安部党委书记、部长赵克志在全国公安厅局长会议上再次强调了公安大数据战略，提出“要深化科技创新，深入推进大数据智能化建设应用，推动大数据建设应用实现新跨越”。

**核心观点**

公安大数据战略仍然是当前公安信息化建设的“核心工程”。自2018年1月公安部成立大数据领导小组后，公安大数据已在多次重要会议或场合被强调。此次在全国公安厅局长会议上又一次提及公安大数据建设，侧面反映出部级领导层面始终重视公安大数据这一战略性工程，公安大数据建设也将不断推进和深化。

公安大数据建设是公安实际“业务”所需。传统公安警务数据中心，无论是在规模上还是在架构上，都不能适应在当前海量数据场景下的数据管理和分析，这将直接对公安形势预判和重大决策造成不利影响。公安大数据平台是通过跨警种、全方位的体系架构，实现囊括从数据汇聚、清洗、存储到数据的处理分析，再到上层的业务应用的统一高效平台。它能够打通公安部门内部的信息壁垒和数据烟囱，并将此前积累的海量数据应用起来，推动公安工作由信息化向智慧化升级转型。此外，打击网络诈骗等新型网络犯罪已成为公安重点工作之一，也迫切需要大数据不断赋能。

公安大数据建设将保持高景气度，预算有望得到优先保障。我们认为公安部门过去在平安城市等领域有着巨大的投入，而未来大数据智能化有望占据核心的位置。大数据作为公安的“一把手”工程，落地后能够向公安各警种赋能，可以在相当长时期内保障各警种实战应用的支撑和服务，落地效果十分明显，因此我们认为相关预算将得到优先保障。目

前公安大数据平台建设正从省部级逐渐向地市级渗透，将维持高景气度。

**投资建议与投资标的**

机构认为公安大数据的建设对于公安打击犯罪、维护社会治安秩序有着重要意义。2021年是建党100周年，我们认为相关公安大数据建设将更早启动。此外，过去建设主要集中在发达省市，今年各地市公安大数据建设也将进入快速普及期，因此相关深度参与到公安大数据建设的企业有望不断受益。

# / 人民日报：套路多损害消费者利益维权难 大数据“杀熟”必须重罚 /

来源 / 人民日报海外版　编辑 / 协会会员处　李苗苗　日期 / 2021-02

同一房型价格不同，同一路线打车却比其他人高出不少，同一家店的外卖吃得总比别人贵……近年来，平台经济快速发展，大数据为消费者提供更具个性的产品和服务。与此同时，网络平台利用大数据“杀熟”也屡见不鲜。日前，中国消费者协会召开网络消费领域算法规制与消费者保护座谈会，指出部分互联网平台利用算法技术进行大数据“杀熟”，损害了消费者的合法权益，必须加强算法应用的有效规制。

**套路多，用户防不胜防**

“师傅，直行就可以吧？”前几天，北京市民赵文军从朝阳区呼家楼打车前往海淀区甘家口北街。打开手机导航软件一看，却发现问题：当时车辆行驶畅通，直行线路并不拥堵，车上导航却显示要绕行，他赶忙要求司机切换了线路。

作为打车软件“重度用户”，赵文军这次没有成为算法“暗箭”的猎物。不过，有些消费者就没这么幸运了。有用户分享自己被“割韭菜”的经历：“我和妻子同时打车，同时发起，同样的起止点，我的定价每次都会贵一些，因为我每天都要打这个路线的车，我妻子是偶尔会打。”

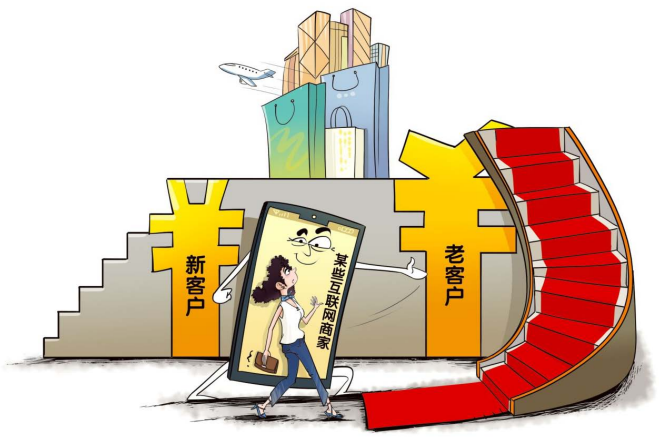
不仅是交通出行，随着平台经济快速发展，网络购物、旅游住宿、电子票务、订餐外卖等生活消费领域或多或少都存在大数据“杀熟”现象。所谓“杀熟”，即同样的商品或服务，老客户看到的价格比新客户要贵出许多。还有消费者明明

是平台会员，但商品价格却比普通用户的要高。

大数据“杀熟”还有更隐蔽的招数。根据消费者投诉，中消协整理出其他类型的算法价格歧视，比如根据消费者所在地区、页面浏览次数等制定不同价格等。

有程序员解释其中“奥妙”：如果平时浏览服装和宠物用品时间比较多，用户会对这些商品价格区间更为熟稔，平台倾向于推荐性价比高的产品。反之，当用户开始浏览平时不怎么看的产品且浏览时间不长，那么算法就会倾向于推荐价格较高的产品。

专家指出，由于不同消费者对价格的敏感度不同，支付意愿也存在差异。所以相比“一视同仁”实行统一定价，差异化的定价行为还能满足商家拉新、提高回购率、提高利润的需求。但是，对时间精力有限的普通用户而言，可谓套路满满，防不胜防。




**维权难，不正之风要刹**

2020年12月发布的《互联网平台竞争与垄断观察报告》显示，消费者对“大数据杀熟”问题反映强烈，认为此举损害消费者利益。中消协指出，任由不公平算法应用无序发展，将损害市场经济的公平有序竞争，还会使消费者面临数据算法压榨，成为技术欺凌的对象。

事实上，中国法律对“大数据杀熟”等算法中的“猫腻”是明令禁止的，相关法律法规也在不断完善中。比如，2020年10月《在线旅游经营服务管理暂行规定》正式施行，其中明确要求在线旅游经营者不得滥用大数据分析等技术手段，基于旅游者消费记录、旅游偏好等设置不公平交易条件等。而此前颁行的《电子商务法》也对“大数据杀熟”作了禁止性规定。此外，有律师指出，根据消费者权益保护法，如果对用户实施价格欺诈，用户可以主张赔偿。

尽管有法律撑腰，但并未刹住平台把老用户当作“韭菜”的势头，维权成本过高而违法成本太低成为关键因素。上海某律所高级合伙人表示，由于工作需要，出差订机票和酒店是家常便饭，被“杀熟”并不鲜见。“顶多跟同事、朋友抱怨几句，几乎没人去真正地维权，一来取证太难，二来需要付出很多时间成本。”

专家指出，大数据杀熟的隐蔽性导致维权举证较难，这也是很多消费者在遭遇此类事件后不了了之的主要原因。据了解，经营者大多以商品型号或配置不同、购买的套餐优惠不同、或者购买时间不同等为借口。即便用户顺利维权，对于市值动辄几十亿甚至上百亿元的平台方而言，罚款可谓不痛不痒。

“连我们这些懂法律的人，都放弃了维权，可见大多数消费者的维权意愿也不会太强，如此一来，就助长了不良风气。”有律师表示。

**强监管，平台不能失序**

更让用户无奈的是，尽管常用的网络平台APP存在种种问题，很多时候却别无选择，因为有些平台已处于垄断地位。

“算法应用不公针对的是社会大众，每个人都可能成为潜在的受害者。”中消协副秘书长栗元广提出，法律应加强算法应用的有效规制。

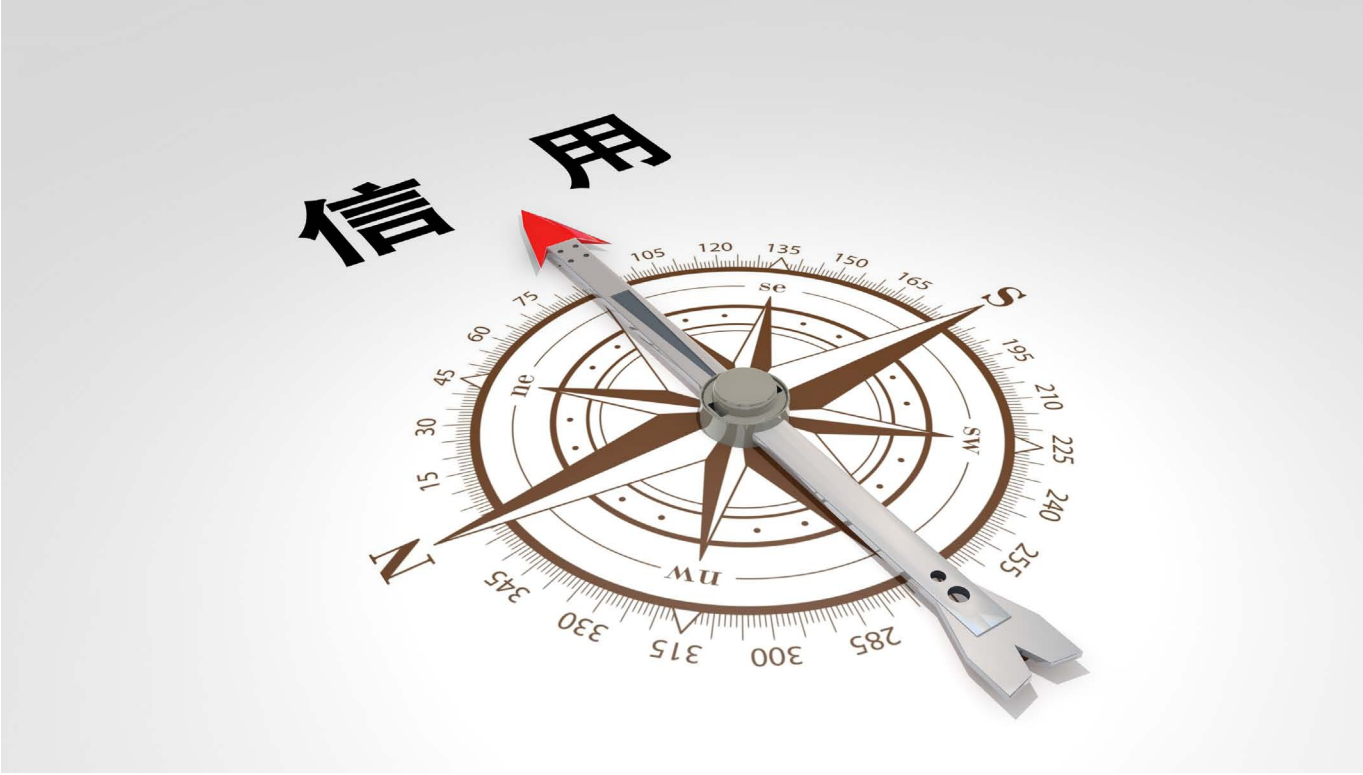
近期，政府对平台经济垄断行为的监管趋于严格。据悉，国家市场监管总局正在制定《关于平台经济领域的反垄断指南》，与线上经济有关的互联网平台都在监管范围之内。与此同时，《反垄断法》修订工作也在积极推动中。下一步，国家市场监管总局将配合行业监管部门共同防范部分企业凭借数据、技术、资本优势造成竞争失序风险，依法加大监管执法力度，强化公平竞争审查。

中消协秘书长朱剑桥表示，社会各界需要共同努力，加强对网络消费领域算法的研究和规制，促进其合理应用，有效维护消费者的合法权益。南开大学法学院教授许光耀认为，消费者遭遇大数据“杀熟”时，对于小额交易而言，提起违约之诉成本较高，也不足以对企业造成威慑。在能够认定行为人拥有支配地位（或达成垄断协议）时，许光耀建议消费者提起反垄断诉讼，对垄断行为的处罚将是行为人上一年度营业额的1%到10%。

对于有违商业伦理的不正当行为，平台和行业不应忽视消费者呼声。中消协呼吁，经营者应当加强自律，守法经营，承担应尽义务和责任。同时，各行业组织应加强引领和督导，使网络经营者合理合规地应用算法，净化行业风气，促进良性发展。

# / 厦门：信用大数据，让守信者处处便利 /

来源 / 新华网　编辑 / 协会会员处　李苗苗　日期 / 2021-02




上医院看病先诊疗后付费，停车先离场后付费，借书免押金，到部分景区、酒店享受优惠……在厦门，随着“白鹭分”应用场景的不断增加，越来越多的市民正实实在在地享受个人信用带来的实惠。

作为国家首批社会信用体系建设示范城市之一，厦门先行先试，在城市信用建设上深耕细作。2018年，厦门市正式发布市民信用“白鹭分”，以厦门市行政管理过程中及市民城市生活中产生的个人信用数据为依据，从基础信息、守信正向、失信违约、信用修复、用信行为五个指标维度设计市民信用评分模型，描绘市民个人信用“画像”、计算市民个人信用分数，并将个人信用分为不良、一般、良好、优秀、极好五个等级。

“我们正致力于开拓更多应用场景，让守信者在生活中处处感受到信用带来的便利，让市民感受到诚信的价值。”“白鹭分”运营单位、厦门信息集团大数据运营有限公司副总经理黄永刚介绍。

**先诊疗后付费　看病迈入“信用时代”**

到医院看病，遇到医保卡内余额不足的情况，患者经常要在诊室和缴费处来回折腾——这是许多患者曾遇到的情况。在厦门，“互联网+信用”融合进一步化解了市民就医流程中尚遗留的痛点。

近日，牙痛不止的厦门市民小方提前预约厦门大学附属中山医院口腔科的门诊号。小方按照预约时间到达口腔科后，

不到十分钟时间，医生就为他完成诊断、开具处方，之后，小方在药房直接取药离院。次日，根据“白鹭分”公众号上的还款提醒，小方通过手机支付320元的医疗费信用账单。

“挺方便的，原来做检查、开药都要分别结算一次，有时候卡里预缴金不够还得充值，现在可以就诊后在手机上操作，一次性搞定。”首次触电信用就医，小方给出了好评。

“先看病后付费”，小方正是凭借“刷信用”，体会到更便捷快速的就医服务。目前，厦门率先在厦门大学附属中山医院和厦门医学院附属口腔医院推行信用就医服务。凡是领取电子健康卡的厦门医保用户，只要达到一定的信用分分值，就能够在医院门急诊中根据相应额度，免除所有缴费手续，诊疗结束后，在约定时间内一次性付清款项。

据统计，自去年6月上线试运行以来，厦门市信用就医服务已累计开通用户20130人，信用就医使用5147人次。

厦门市卫生健康委相关负责人表示，信用就医服务将社会信用引入医疗领域，有利于强化诚实守信的社会价值。下一步厦门将增加信用就医试点医院，计划将信用就医延伸到住院环节，扩大信用就医的内涵和服务范围。

**停车先离场后缴费　市民出行更通畅**

“你好，我是白鹭分用户。”在海沧区上班的吴颖和同事驾车外出用餐，将车辆停在亚桥花园路面停车位后，她对管理车位的收费员“亮”出自己的白鹭分用户身份。

“一般停车后，收费员会先来收取一定的停车费，离场的时候还会拦住车主，补缴剩下的停车费。”吴颖说，自从海沧区推出信用停车服务，她就在白鹭分公众号上绑定自己车牌，开通微信免密支付，省去了不少麻烦。

约半个小时后，吴颖和同事直接驾车离开，不到一分钟，她就接收到缴费的微信通知，显示她的车辆在12:02进场，12:35出场，使用白鹭分享受9折优惠，系统已自动扣除停车费4.50元。

2018年底，由厦门信息集团大数据运营有限公司打造的白鹭分信用停车服务在海沧区率先推出，市民的白鹭分达到530分以上即可绑定车牌，开通微信免密支付，享受信用停车服务。

“我的白鹭分有700分，白鹭分达到680分就能享受9折优惠。”吴颖说。

据了解，厦门白鹭分信用停车服务范围不断扩大，今年上半年在思明区上线，不久前，该服务延伸到湖里区的市政车位，实现厦门岛内市政车位全覆盖。截至目前，信用停车共覆盖厦门全市超过8000个路边停车位。

“市民省去了等待收费的时间，出行更加方便快捷。”黄永刚说，白鹭分“信用停车”覆盖范围还将进一步扩大到厦门其他区。

**越守信越有获得感　鼓励市民参与信用建设**

如今，厦门白鹭分注册用户总数已突破118万，市民自愿使用量（查分次数）超过897万次。白鹭分覆盖了包括图书馆“免押借阅”、公共自行车免押骑行、信用停车、信用就医等“信易+”应用场景40个。

截至目前，厦门惠民信用停车福利时长已突破72万分钟，免押借阅量超过600万册，信用就医降低患者就诊时长突破102940分钟。更多信用场景还在开发中——例如，通过与一些区行政服务中心对接“白鹭分”信用应用服务，实现信用分值高的市民免排队优先办理相关业务，选择信用容缺办理等。

值得关注的是，作为个人信用身份证的“白鹭分”，还将记录市民对城市作出的贡献——市民不仅可以通过坚持信用习惯、多使用“白鹭分”应用来实现分数的增长，还可通过积极参加志愿者服务、无偿献血等公益服务来提升信用。

基于市民个人信用数据，厦门信用建设在更多与民生息息相关的领域得到体现。据了解，近两年，厦门推进家政从业人员诚信信息平台建设，打造家政服务消费溯源体系，正逐步形成全市家政服务业大数据信息的集成共享。

需要雇佣家政阿姨的市民在“白鹭分”微信公众号上点击进入“鹭岛家政”应用场景。该应用链接了厦门市征信平台、公安背审系统、厦门市民个人信用分（白鹭分）系统等多个市级系统。雇主（消费者）只需要通过“鹭岛家政”小程序扫描家政服务人员的“家信卡”，就可以查看其从业记录、培训记录以及信用记录。还可以在雇请家政人员后填写相关服务记录，对已服务家政人员进行评价和投诉，由此实现家政人员信息“可查询、可追溯、可评价”。

让信誉建设全覆盖、无死角，厦门正形成更加立体化、网格化的信誉评价体系，培育良好的社会守信氛围，让守信市民越来越有获得感。

# / 税收大数据显示：2020年企业销售收入同比增长6% /

来源 / 新华社　编辑 / 协会会员处　李苗苗　日期 / 2021-02

增值税发票等税收大数据具有时效性强、覆盖面广、颗粒度细的优势，能够较为迅速全面准确客观反映经济运行情况。在20日举行的国家税务总局新闻发布会上，税务总局收入规划核算司司长蔡自力说，增值税发票数据显示，2020年，全国企业销售收入同比增长6%，比前三季度提高4.2个百分点。其中四季度增长15.7%，比三季度提高3.3个百分点，反映出我国经济市场活力和发展潜力在快速释放。

据蔡自力介绍，2020年，随着统筹疫情防控和经济社会发展政策落地生效，我国经济运行稳步复苏回升。主要呈现六个特点：

**一是工业企业销售稳步回升，装备制造业支撑作用增强。**数据显示，2020年，工业企业销售收入同比增长2.7%，在10月份累计增速转正后继续稳步回升。其中，四季度增长11.1%，比三季度提高1.9个百分点。工业41个大类中，有26个实现正增长。制造业复苏较快，全年销售收入同比增长3.2%，特别是装备制造业支撑作用持续增强，同比增长7.1%。

**二是商品消费恢复较快，线上消费保持快速增长。**数据显示，2020年，批发零售业销售收入同比增长5%，其中四季度增长15.4%，比三季度提高4.1个百分点。疫情改变了居民消费方式，“非接触式”消费快速发展，线上线下平台加快融合，互联网零售、外卖送餐服务同比分别增长29.2%、31.6%。

**三是投资相关行业增长迅速，新基建发展提速。**数据显示，2020年，与基建投资相关的土木工程建筑销售收入同比增长19.7%。特别是5G基站、特高压、城市轨道交通、环保等新基建领域投资增长较快，带动通讯和电力输送设施工程、城市轨道交通工程、节能工程建筑销售收入同比分别增长10%、13.4%、20.8%。

**四是交通物流业稳步恢复，道路货物运输等增长较快。**数据显示，2020年，交通物流业销售收入同比增长9.6%，高于全国总体水平3.6个百分点，其中四季度增长22.8%，比三季度提高10.3个百分点。特别是道路货物运输、仓储、邮政物流业，销售收入同比分别增长9.9%、16.5%、23.1%。

**五是能源结构逐步优化，环保相关产业加速发展。**2020年，水力、核力、风力等新能源发电行业销售收入占发电行业总体的比重为37.9%，比2019年提高3个百分点。其中，太阳能发电、风力发电销售收入分别增长16.5%、14.1%。资源综合利用、生态保护环境治理销售收入同比分别增长17.8%、15%，反映出环保产业持续加速发展，绿色理念更加深入人心。

**六是民营经济领跑优势明显，私营企业增长较快。**数据显示，2020年，民营经济销售收入同比增长8%，高于全国企业总体水平2个百分点，其中四季度增长18%，比三季度提高3.2个百分点。私营企业销售收入同比增长9.6%，高于民营经济总体水平1.6个百分点。

# / 金融科技企业50报告：AI、大数据等底层技术企业大幅增加 /

来源 / 经济观察网　编辑 / 协会会员处　李苗苗　日期 / 2021-02

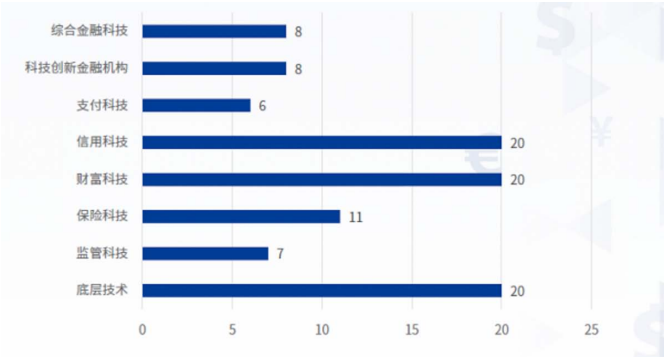
1月18日，毕马威发布《2020中国领先金融科技企业50》报告及评选入围榜单。本届榜单围绕科技与数据、创新与变革、金融服务普及、资本市场认可度、发展前瞻度五大维度对企业进行量化评估。

报告指出，新冠疫情爆发和疫情防控进入常态化之后，各类无接触金融服务需求的大幅提升、数字化转型需求迫在眉睫，金融企业纷纷通过布局金融科技子公司、设立金融科技专项基金等方式加大了对金融科技的投入与创新应用。

此外，疫情也倒逼各类资管机构加快运营的数字化转型。资管机构运用RPA智能机器人、大数据、人工智能等技术打造一体化智能运营能力，显著提升企业运营效率，降低运营成本。

从专业领域分布来看，入围企业集中分布在信用科技、财富科技、底层技术这三个领域。

专业领域分布：

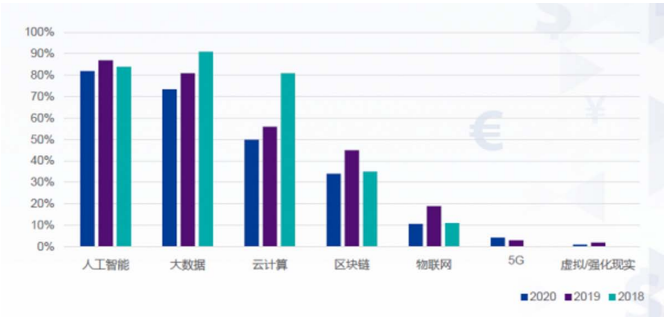



从分布地域来看，入围企业所在地前三位的分布是北京（34家）、深圳（25家）、上海（23家）。

地域分布：

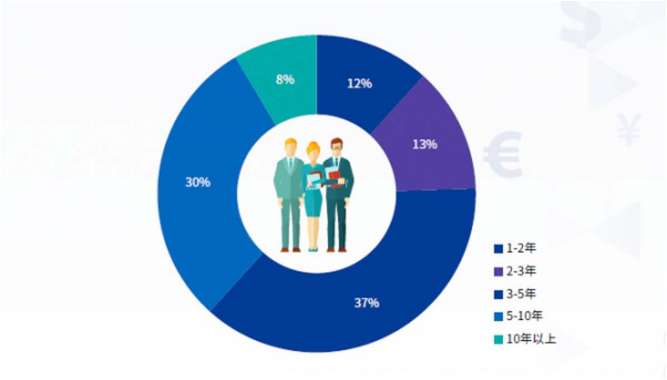
从主要技术要素分布来看，企业最主要集中在的是人工智能、大数据、云计算。

主要技术要素分布：

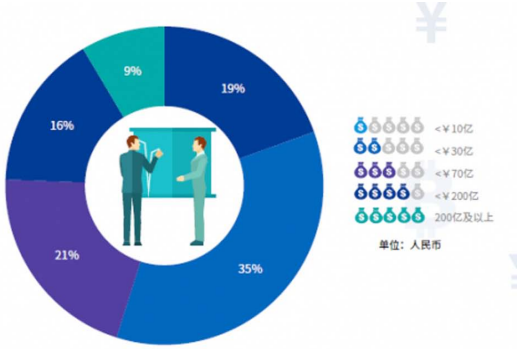



值得注意的是，入围企业中成立时间10年以上的仅有8%，成立3-5年的占比37%，成立5-10年的企业占比30%。

成立时间分布：




而从估值来看，35%的企业估值少于10亿元，市值超过200亿元的企业占比为9%。




毕马威中国亚太区及中国主席陶匡淳表示，在评选过程中看到中国金融科技市场的巨大潜力，上榜的100家中有50家屡次入选，还有50家是新上榜的企业，这与中国金融科技产

业持续发展趋势相符合。

毕马威中国金融科技合伙人(华南区）蔡正轩表示，此次疫情可以说是对各行各业的一次全链路压测，其中会暴露出很多商业模式的问题、组织效率的问题、灾备能力的不足。所有金融行业和金融科技的从业者，如果能前瞻性地应对风险和及时调整，可以使风险管理和精细化运营水平与效率提升到新层次。

对于金融科技的底层技术如何与业务深度融合，蔡正轩表示，底层技术越来越强调普惠性智能，底层技术方向朝着规模化、低成本、普适性发展；下游金融应用场景会不断拓宽，用户管理、产品定价、渠道营销、运营模式等业务场景将产生重大变革。

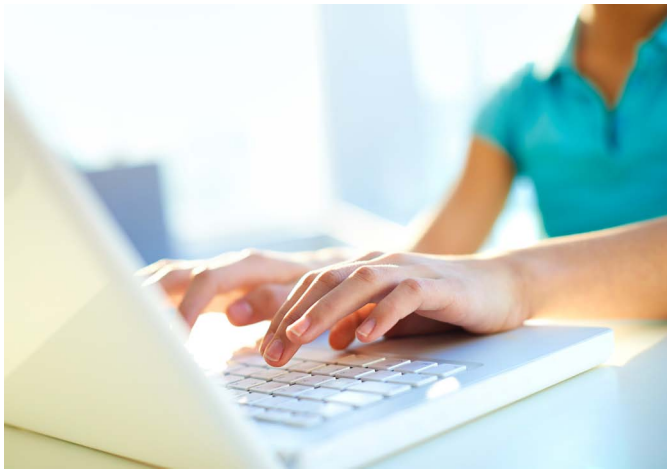
毕马威中国金融业务税务主管合伙人张豪表示，金融科技行业的发展需要底层技术创新的配合，最近两年可以看到，投资机构、投资人、企业都在将关注点从场景创新转向了底层技术创新。

蔡正轩也表示，金融科技与金融细分领域结合，带来了新的业务形态；在金融科技背景下，各类科技平台、互联网巨头、金融行业头部企业纷纷部署了综合金融科技类企业，争相抢占、开拓新市场，并打造全新的金融科技生态圈，每个细分领域及环节都可能出现独角兽。

随着技术迭代不断加快，金融科技领域也出现了数据割裂的问题，也就是数据大量沉淀在大企业手中，而真正做底层技术创新的则是那些相对体量更小的、初创的科技公司。

张豪认为，解决数据割裂问题需要从两个方面同时进行，一方面是开放数据，另一方面是通过技术增强数据的安全性。

“数据开放就是对基础数据实现共享，比如开放式银行，虽然推进有困难，但这是方向；对于数据的保密性，其实现在的技术已经可以通过暗码对数据进行解析，并保护用户的隐私安全。”张豪表示，金融科技将会让真正拥有底层技术的企业获得更大的市场和发展空间。



# / 2021年大数据分析的5大挑战 /

来源 / 企业网D1Net 编辑 / 协会会员处 李苗苗 日期 / 2021-02

2021年已经到来，现在是深入研究大数据分析面临的挑战的时候了，需要调查其根本原因，本文重点介绍了解决这些问题的潜在解决方案。

在开始使用大数据分析系统时，组织最好考虑周全。因为一旦大数据分析已经启动运行，任何修复都可能成本高昂。

在当今的数字世界中，很多组织采用大数据分析系统改进业务决策、增强责任感、提高生产率、做出更好的预测、监控绩效，并获得竞争优势。然而，许多组织在战略层面上使用商业智能分析存在一些问题。根据调研机构Gartner公司的调查，87%的组织表示商业智能(BI)和分析成熟度较低，缺乏数据指导和支持。业务数据分析面临的问题不仅与分析本身有关，还可能与深层次的系统或基础设施问题有关。

**1.解决方案无法提供新见解或及时的见解**

很多组织投资采用新的大数据业务分析解决方案，力求获得独特的见解，以帮助领导者做出更明智的业务决策。但有时，新系统提供的见解似乎还不如以往采用的系统提供的见解水平和质量。组织可以从业务或技术的角度来解决这个问题。

(1)数据不足

有些组织可能由于分析数据不足，无法生成新的见解。这可能是由于缺乏数据集成或数据组织不当造成的。

在这种情况下，可以进行数据审核，并确保现有数据集成提供所需的见解。新数据源的集成也可以消除数据的缺乏。还需要检查原始数据是如何进入系统的，并确保所有可能的维度和指标均已经公开并进行分析。最后，数据存储的多样性也可能是一个问题。可以通过引入数据湖来解决这一问题。

(2)数据响应慢

当组织需要实时接收见解时，通常会发生这种情况，但是其系统是为批处理而设计的。因此有些数据现在仍无法使用，因为它们仍在收集或预处理中。

检查组织的ETL(提取、转换、加载)是否能够根据更频繁的计划来处理数据。在某些情况下，批处理驱动的解决方案可以将计划调整提高两倍。另一个选择是使用一种称为Lambda架构的方法，该方法允许组织将传统的批处理管道与快速的实时流结合起来。

(3)新系统采用旧方法

虽然组织采用了新系统。但是通过原有的办法很难获得更好的答案。这主要是一个业务问题，并且针对这一问题的解决方案因情况而异。最好的方法是咨询行业专家，行业专家在分析方法方面拥有丰富经验，并且了解其业务领域。

**2.不准确的分析**

对组织来说，没有什么比不正确的分析更糟糕的事情了，这个问题需要尽快解决。

(1)源数据质量差

如果组织的系统依赖于有缺陷、错误或不完整的数据，那么获得的结果将会很糟糕。数据质量管理和涵盖ETL过程每个阶段的强制性数据验证过程，可以帮助确保不同级别(语法、语义、业务等)的传入数据的质量。它使组织能够识别并清除错误，并确保对某个区域的修改立即显示出来，从而使数据纯净而准确。

(2)与数据流有关的系统缺陷

当由于开发、测试或验证过程中的人为错误而忽略或没有完全满足系统的需求时，就会发生这种情况。

通过对开发生命周期进行高质量的测试和验证，可以减少此类问题的发生，从而最大程度地减少数据处理问题。即使使用高质量数据，组织的分析也可能会提供不准确的结果。在这种情况下，有必要对系统进行详细检查，并检查数据处理算法的实施是否无故障。

**3.在复杂的环境中使用数据分析**

这个问题可能会使为创建有效解决方案而投入的所有努力付诸东流。如果使用的数据分析变得太复杂，可能会发现很难从数据中提取价值。复杂性问题通常归结为用户体验(当用户很难浏览系统并从报告中获取信息时)或技术方面(当系统设计过度时)。

(1)数据可视化显示凌乱

如果组织的报告复杂程度太高。这很耗时或很难找到必要的信息。可以通过聘请用户界面(UI)/用户体验(UX)专家来解决此问题，这将帮助组织创建引人注目的用户界面，该界面易于浏览和使用。

(2)系统设计过度

数据分析系统处理的场景很多，并且为组织提供了比其需要还要多的功能，从而模糊了重点。这也会消耗更多的硬件资源，并增加成本。因此，用户只能使用部分功能，其他的一些功能有些浪费，并且其解决方案过于复杂。

确定多余的功能对于组织很重要。使组织的团队定义关键指标：希望可以准确地测量和分析什么，经常使用哪些功能以及关注点是什么。然后摒弃所有不必要的功能。让业务领域的专家来帮助组织进行数据分析也是一个很好的选择。

**4.系统响应时间长**

即使输入数据已经可用，数据分析系统仍然花费大量时间来分析数据，并且需要快速提交报告。对于批处理而言，它可能并不那么重要，但是对于实时系统而言，这种延迟可能会花费很多成本。

(1)数据组织效率低下

也许组织的数据组织起来非常困难。最好检查其数据仓库是否根据所需的用例和方案进行设计。如果不是这样，重新设计肯定会有所帮助。

(2)大数据分析基础设施和资源利用问题

问题可能出在系统本身，这意味着它已达到其可扩展性极限，也可能是组织的硬件基础设施不再足够。

这里最简单的解决方案是升级，即为系统添加更多计算资源。只要它能在可承受的预算范围内帮助改善系统响应，并且只要资源得到合理利用就很好。从战略角度来看，更明智的方法是将系统拆分为单独的组件，并对其进行独立扩展。但是需要记住的是，这可能需要对系统重新设计并进行额外的投资。

**5.维护成本昂贵**

任何系统都需要对其维护和对基础设施进行持续投资。每个组织都希望尽量减少这些投资。因此，即使对降低维护和基础设施的成本感到满意，也需要重新研究一下系统，并确保没有支付过高的费用。

(1)过时的技术

每天都会出现可以更快、成本更低处理更多数据量的新技术。因此，与现代技术相比，组织的分析技术迟早会过时，需要采用更多的硬件资源，并且维护成本会更高。也难以找到愿意基于传统技术开发和支持解决方案的专家。

组织最好的解决办法是采用新技术。从长远来看，它们不仅可以降低系统的维护成本，还可以提高可靠性、可用性和可扩展性。逐步进行系统重新设计，并逐步采用新元素替换旧元素也很重要。

(2)并非最佳的基础设施

基础设施总有一些优化成本的空间。如果组织仍然采用的是内部部署设施，将业务迁移到云平台可能是一个不错的选择。使用云计算解决方案，组织可以按需付费，从而显著降低成本。如果对安全性有任何限制，仍然可以迁移到私有云。如果已经在云平台上运营，需要检查是否有效地使用，并确保已经实施了所有的最佳实践来削减开支。

(3)选择了设计过度的系统

如果组织没有使用大多数系统功能，则需要继续为其使用的基础设施支付费用。组织根据自己的需求修改业务指标并优化系统。可以采用更加符合业务需求的简单版本替换某些组件。

**结论**

组织可以调整现有的业务分析平台，但可能会变成一项艰巨的任务。如果组织在新的解决方案设计和实施中错过了某些内容，则可能会浪费时间和资金。

# / 数据产品经理必看！数据看板搭建攻略 /

编辑 / 协会会员处 李苗苗 日期 / 2021-02

在后台中，数据面板能够直观反映出业务变化，并有助于决策层发出业务调整与决策。那么搭建数据看板的时候，要如何搭建呢？又有哪些注意要点呢？

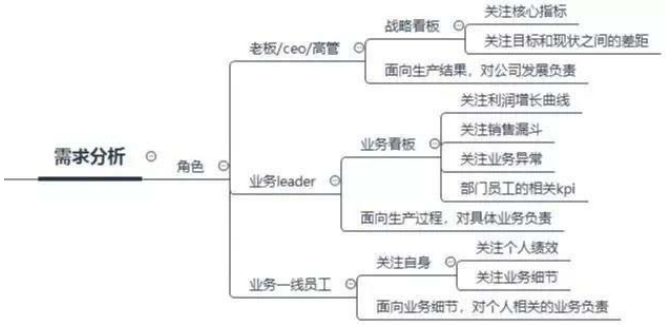
## 01
## 产品背景分析

01 - 什么是数据看板？

数据看板一般用作后台系统的首页，主要呈现公司当前业务相关或运营管理相关数据和图表，方便公司内部人员实时了解公司内情，掌握业务发展情况，并能够对数据变化做出业务决策。

02 - 谁在使用数据看板？

数据看板是公司内所有人员都会关注的，无论是公司高层还是基层业务操作者都会关注相关的数据指标。只是不同的角色的需求不同，如下图：




不同的角色的默认看板页面是不一样的，要设计多份看板页面，至于不同页面是否可以互通，这就要取决于公司业务数据是否属于商业机密，牵扯到RBAC权限模型的设计，这里就不多赘述了。

## 02
## 设计看板的前置条件：报表设计

设计看板可不是脑子一热就开始画原型了，那样得到的看板是不切合实际的，而且也不一定准确，尤其是对于数据大量且复杂的业务而言，需要先设计好报表。数据看板中有不少数据是需要后台去实时计算的，在数据量太大的情况下，筛选条件和统计项较为复杂时，图表加载速度会变慢，也会耗费大量系统资源。而如果先设计好数据报表，在数据看板中，查看图表的速度才不会收到影响。

不要急于数据可视化的设计，需要数据报表所呈现的数据经过多次测试没有问题，再经过至少一轮的迭代优化，再着手进行可视化方案的设计。

数据报表的设计需要经过以下流程：

01 - 构建分析体系

· 面向业务结果的分析

b端产品经理不同于c端，业务是一切产品设计的核心，必须围绕业务进行产品设计。首先要明确分析目的，并进行业务诊断，最重要的是打通分析链路。

例如要分析利润曲线，与利润相关的数据有利润率、在架率、产品种类、报价策略、市场行情变动等

根据公司具体业务不同，明晰哪些是需要重点关注的，哪些是可以暂时搁置的，特别是要跟业务团队多沟通协调，共同商议出一个分析框架。

一般可采用枚举法，列举可能干扰到核心业务数据所有因素，再一一排查。

· 面向生产过程的分析

公司不仅要求分析业务结果等业务强相关数据，生产过程关系到每个人的工作效率和工作成果是否达标，生产过程同时也是对人员kpi的考核。

如对销售的拜访客户数、销售线索获取数量、签单客户数，贡献利润占比等，对客服的工单处理数量、客户评分、分钟会话数量等，该部分分析主要根据使用者的操作记录产生数据。

02 - 定义观察指标

· 进行维度下钻

定义好分析框架后，要进一步确定观察指标，主要是对观察维度进行拆分，逐渐细化，只有在更精细的维度分析数据才最有可能得到准确的答案。

比如在架率又可细分为日间在架率和夜间在架率，还可继续细化为某某产品在架率，甚至可以细化到业务员王二在日间12:33误操作某交易量极大的产品，造成利润损失。这样就打通了从业务结果——生产过程的分析链条。

又比如报价策略有可细化为具体sku的报价问题，该数据又可关联到上游供应商近期内价格调整频繁，而报价策略未结合该特点导致该产品报价过高/过低。

维度细化的过程中，可能会逐渐拆分为一级指标、二级指标……n级指标，但并非拆分中的每一个维度都要作为观察指标。是否作为报表使用要结合公司业务特点并且和业务部门多次沟通协调达成共识。

在维度下钻的过程中确定哪些维度是需要进行统计的，哪些是不需要用到的。

· 统一数据口径

数据口径指的是对于数据的定义，如果系统内部统计口径不一致，就会导致业务分析出错。如果不对数据口径进行清晰的定义，开发人员在调用数据时可能会直接采用已有的口径，这个已有统计口径不一定符合当前数据指标的需要。

产品人员应该单独书写一份关于数据口径的定义的文档，并上传至系统内部的文档共享系统，方便业务人员随时查阅。如果系统允许，尽可能与业务部门达成共识，形成一套标准的数据口径。

· 定义统计频次

确定好需要统计的数据维度形成观察指标后，要对观察指标分类处理，有些指标属于关键性因素，有些指标属于次要因素。有些指标单位时间内产生的数据波动较大，统计频次可能要到分钟级。而有些指标数据波动较小，可能一天只有几次变动，统计频次可设定为小时级。

03 - 设计呈现形式

一开始的呈现形式以简单、好调整为主，不必急于线上化，也不必追求酷炫的交互效果。严格遵循mvp原则，用最小功能集合去迭代产品。

报表在设计初期必然面临指标的频繁调整，以及数据不准确，统计范围需要变更等种种问题。初期可采用系统定时通过邮件发送报表，报表的形式可以是一张简单的excel表格，这样先试用一段时间，确定要分析的数据字段、监控指标等没有问题后，再着手进行线上化的设计。

确认好线下的报表迭代优化完毕后，设计线上化的方案，一般考虑以下几个方面：

· 默认查询时间，统计具有延时的特点，有些数据统计必须等待订单完成才能纳入统计范围，比如默认查询时间定为t-6至t-2

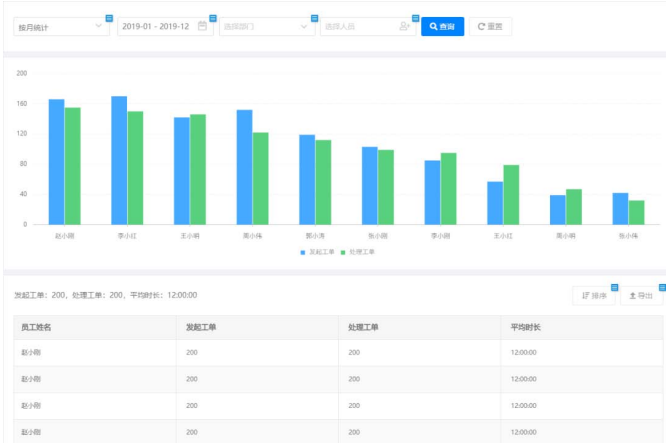
· 排序规则，默认按照哪一项数据排序

· 筛选项设置，哪些数据选项需要筛选

· 统计项设置，哪些数据需要合并统计

· 其他极端情况考虑，数据极限值，最大时间范畴，取数为空时的处理等等

大致的形式如下，但在实际业务中数据远比这个复杂：




04 - 复盘报表设计

报表线上化固然有其优点，方便、快捷，效率高。但是也存在一些弊端，比如不够灵活，excel表格可以方便做成各种你想要的数据透视表，而线上只能固定其中的几种。线下还

可以采用任何关联要素利用vlookup进行匹配，线上则已经固化了分析指标。

因此，产品上线一段时间后，要及时复盘功能是否好用，是否得到认可。

· 使用率：可通过数据埋点看该项功能的使用率，使用率不够要和业务部门沟通问题，及时迭代优化

· 观察是否实现最初想要得到的目的，分析链路上是否还存在障碍

· 用户满意度: 与业务部门沟通，看使用过程中有什么不方便的地方，效率方面是否还可以得到提升

## 03
## 数据看板的设计

当数据报表完善之后，那么实现数据可视化也就非常方便了，效率是b端产品的核心，设计看板也不应该花费太多时间在交互体验上。这里主要满足以下几个设计要点：

· 简单高效，优先满足查询效率，而不是酷炫的交互

· 信息具有强关联性而不是孤立的一个数据，具体就是要有环比、同比来体现变化

· 数据图表的刷新频次和统计频次要符合业务的需求，最好能做到实时更新

· 选用的数据能够体现出趋势和规律，对于无趋势特性的数据，直接展示数字比较好

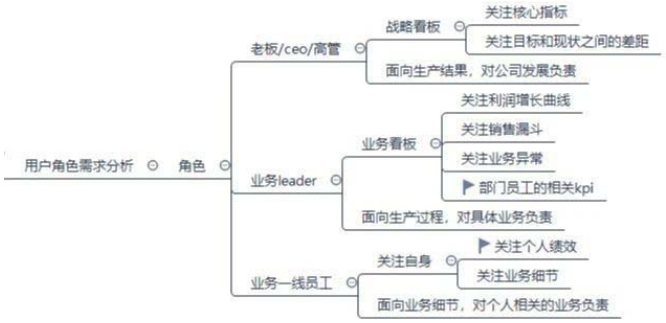
· 对于不同的数据指标，不同的数据特性需要选用合适的图表。

· 数据波动、对比、排序，不同的衡量方式也应该选择其对应的图表类型

01 - 需求分析

数据报表可能有十几种报表，几十个数据指标，但这些并不是都需要呈现出来的，选用哪些数据指标和衡量方式，需要和业务团队沟通确定。

需求是分层级的，有些需求处于核心业务的需求，有些属于经营管理层面的需求，而前者永远是公司发展的核心，因此其重要性和优先级要高于后者。




对于上图来说，部门员工的相关kpi，个人绩效属于经营管理层面的数据要求，第一期工程可以先不考虑实现，先确保实现核心业务相关的数据看板的可用性和准确性。

02 - 战略看板设计

战略看板设计主要突出简洁、唯结果论的特点，要主次分明，并且多的展现纯数字内容而非图表，下图举例：

一般设置6个左右的数据指标，有利于专注于分析最重要的指标。

需要支持自定义配置，可根据业务发展需要灵活调整看板需要呈现的数据。

之所以采用数字形式而不使用各式各样的图表，一是为了保证第一时间掌握业务现状，二是公司中高层对业务数据非常敏感，不需要曲线来表示规律，只用看到数字就能看出业务异常。

产品经理也应该培养数据敏感性，以便于能够理解业务，做出更符合业务需求的数据产品。

03 - 业务看板设计

业务看板设计主要突出垂直细分的特点，研究业务增长就是业务增长的图表，研究活跃用户就是活跃用户的看板，或许这其中会有写关联性，但这种关联是弱关联，如果数据看板之间含混不清，那么分析业务也会有很多阻碍。

业务看板与战略看板最大的不同点是，注重生产过程，所以多使用增长曲线来体现数据规律和数据变化。

研究增长曲线：

销售漏斗：

上图所示仅为单一维度，实际业务场景往往是一个关键业务指标关联着其他的数据指标，图表要支持维度下钻，点击能够查看二级指标的图表信息。

要能够支持x轴平移和缩放，y轴的刻度能够根据数据大小自适应。

业务看板解决的核心问题是：及时发现业务问题和便于分析业务异常，视觉设计上要注重各个指标的色彩鲜明，便于察觉指标与指标间的关系。

业务看板描绘的业务不同，数据不同，需要安排在不同的页面上，业务看板需要足够细分，绝对不能互相杂糅，纠缠不清。

04 - 选用合适的图表类型

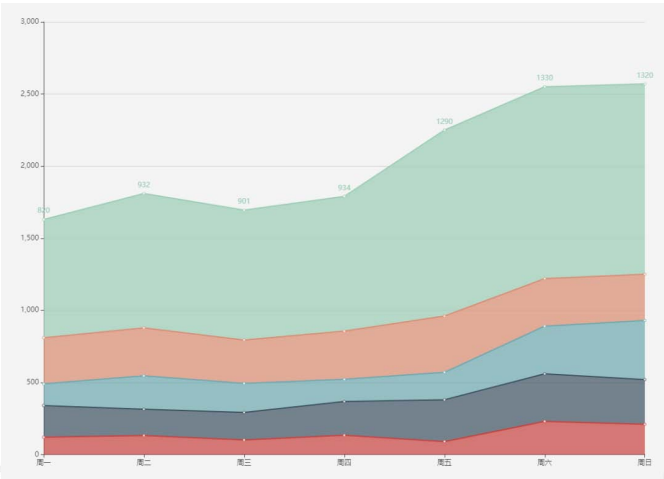
选择合适的图表类型能够最大程度满足视觉可视化的呈现效果，满足业务数据查看和分析的需求。

以下选用几个典型的业务场景做说明：

（1）描述业务增长曲线，利润曲线、销量曲线、用户数量等，重点表示业务发展的趋势

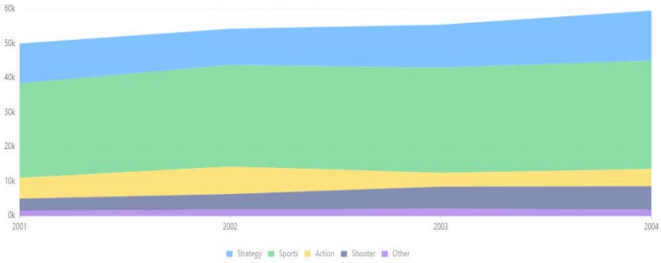
常规面积图：

所有数据都从相同的零轴开始，使用与展示总量与分量之间的关系，必须设置透明度以避免重叠

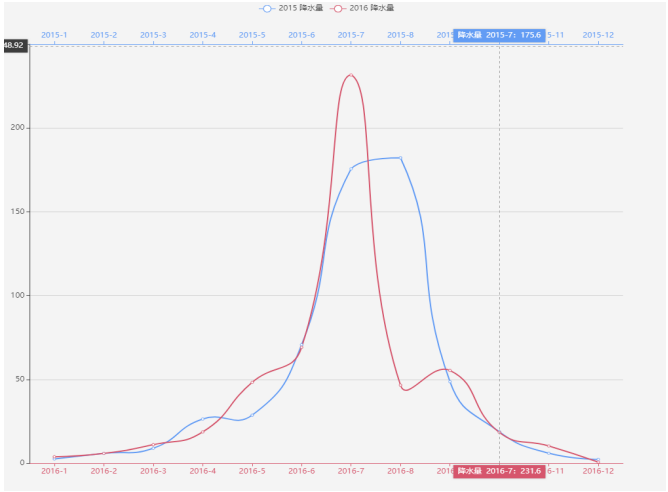


堆叠面积图：

数据起点为上一个数据集，主要用于展示数据分量占数据总量的百分比，然后对分量所占百分比进行对比



折线图：




折线图和面积图都可以用来展示连续数据随时间变化的趋势，但选用还是有其侧重点：

· 常规面积图适合表现总量和分量之间的关系，比如总量是总用户量，而分量是各个渠道的用户，那么选用面积图较为合适

· 堆叠面积图适合表现分量所占数据总量的比例

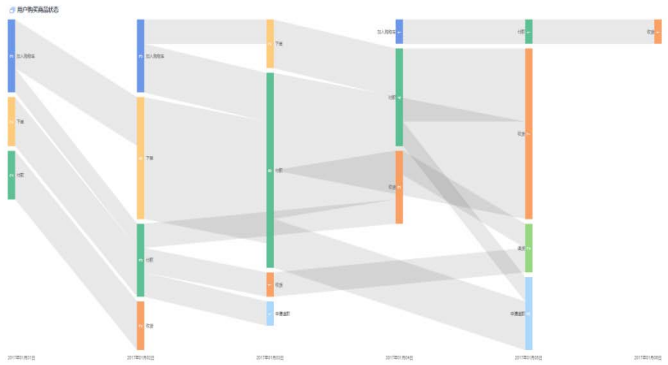
· 折线图更适用于展现分量与分量之间的对比（不含总量），避免了面积图阻塞的问题

· 通常情况下，需要展示数据总量的业务，使用面积图会更好些

（2）描述业务流程问题

描述业务流程最常用的图表是漏斗图，销售漏斗图、用户留存漏斗图，但漏斗图只能表示单一流程，如果想要表示具体的用户数据流向，和多维度更加复杂的流程数据，就很难表示了。

桑基图：

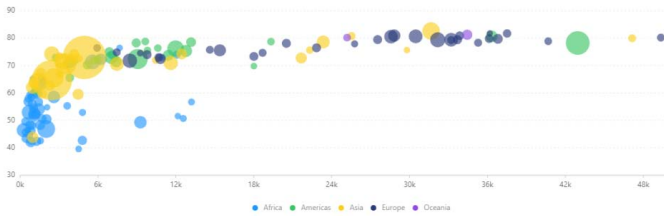


桑基图能够很好地表示用户从加入购物车到收获的全流程，主支宽度总和与分支宽度总和相等，不同宽度代表不同流量的大小。

漏斗图这里就不过多赘述了，比较简单。

（3）描述数据分布

气泡图：




气泡图最适合展示数据的分布范围，横坐标可设置为时间，纵坐标可设置为商品的价格，表明商品价格区间随时间变化的趋势，而不同的颜色表示不同的商品，这样对于价格频繁变动的商品，查看其每天的价格分布以及趋势很有帮助。

不常用以及过于简单的图表，这里就不再详细展开了，使用图表以好用、够用为原则，不必过于追求复杂图表，否则业务团队觉得麻烦，不愿意使用，那就背离了产品设计的初衷了。

另外，自行开发可视化图表耗时耗力，常见的解决方案是使用ECHARTS 、QCHART、AntV等开源可视化库，只需调用相应的api接口即可。

## 04
## 其他注意事项

由于b端业务的特性，后台系统不对外开放，仅企业内部人员使用，因此着重强调的还是围绕业务问题进行行业务数据建模，打通分析链路、维度下钻和报表设计等层面，看板的设计反而较为简单（b端对于视觉体验的要求不高），况且市面上已经有很多非常成熟的可视化解决方案可供调用，视觉体验的设计下限比较高，因此后台产品经理应该把主要精力集中在后端逻辑层面。

产品经理要尽可能的多参与前期业务设计的过程中去，旁听业务讨论也有助于理解业务部门痛点和难点在哪里。业务调研和访谈的形式不够深度，业务方在表达诉求时也可能会偏离他们的初衷，产品经理只有深入业务逻辑才能对业务方提出的需求多加辨别分析，形成有产品经理专业意见的设计方案。

# / 战"疫"启示录：追本溯源，求解智慧城市建设最大公约数 /

来源 / 中国IDC圈　编辑 / 协会会员处　李苗苗　日期 / 2021-02

智慧城市的建设通常是定位在城市管理的信息化亮点上，通过本次新冠肺炎疫情，我们发现无论是智慧城市的建设理念、信息化顶层设计，还是云计算、大数据、物联网、5G、区块链等各种新技术的应用，其本质首先应是服务于城市的生存，以此为载体应用于产业发展。

新冠肺炎疫情期间，在全民抗击疫情的同时，一些地方政府、智慧城市领域专家也在反思过去10余年我国智慧城市的建设，提出了很多良策和建议。2月10日，上海市人民政府官网发布了《关于进一步加快智慧城市建设的若干意见》（以下简称《意见》）。《意见》聚焦政务服务“一网通办”、城市运行“一网统管”、全面赋能数字经济三大建设重点，夯实“城市大脑”、信息设施、网络安全三大基础保障，加快推进新一轮智慧城市示范引领、全面建设，不断增强城市吸引力、创造力、竞争力。新冠肺炎疫情期间，我国10余年的智慧城市建设和累计数千亿规模的投资并没有起到预期的作用，与十几年前的非典相比，疫情防控的数据通过移动互联网是相对透明的，但其他的防控方式和方法相对于十几年前并没有太大变化，因此，笔者认为在这个节点上应反思一下智慧城市的初始需求，我们的初衷是什么，我们的建设路径是什么。

**城市建设的需求层次理论**

对于人的需求层次，亚伯拉罕·马斯洛（Abraham Harold Maslow）于1943年提出“马斯洛需求层次理论”，其基本内容是将人的需求从低到高依次分为生理需求、安全需求、社交需求、尊重需求和自我实现需求，为心理学研究、企业人力资源管理提供了理论支持。

在笔者看来，从某种意义上来说，人的第一位需求应该是“安全需求”而不是“生理需求”。人类首先要解决的是生存、安全问题，如果生存都没有解决，诸如食物、睡眠等更深层次的生理需求也无法实现。城市是人类的生活载体和活动集合，城市首先要解决市民的生存问题，诸如大规模传染病、城市较大内涝、地震、台风、食品安全、饮用水安全、空气化学污染等。

每一种人为或者自然灾害来临之前我们都要保障城市的生存——市民人身安全、市民财产安全、公共设施安全等“安全需求”，这是第一层。第二层才是城市的清洁卫生、社会治安、生活便捷、城市噪声污染、雾霾治理等“生理需求”。第三层是社会交往需求，比如各类社区交往活动、公共文体活动、论坛与展会等“社交需求”。第四层“尊重的需求”是城市建设的各方参与者都能平等地获得信息、充分表达利益诉求和民意、社会公平正义、机会均等诸如此类。第五层“自我实现的需求”是能够提供各种渠道给城市参与者实现自我价值，达到城市的管理者、市民、企业的发展与城市的发展相得益彰，如志愿者渠道、行业联盟等。

**智慧城市建设的需求层次理论**

2008年，IBM提出了“智慧地球”理念，意在通过“互联网+物联网”为各行各业提供便捷的解决方案，但没有提出要基于城市安全。智慧城市建设需求层次理论属于城市建设的需求层次理论的子集，因此也应该分为五大层次，如图1所示。

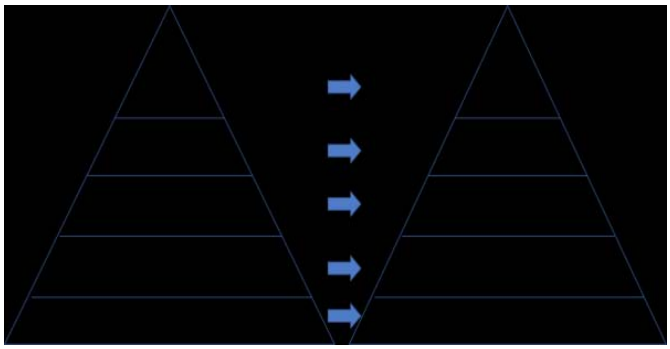


图1 城市的建设需求层次与智慧城市的需求层次

从基本的层面来看，城市发展起于农业经济时代的商业活动，兴盛于工业化时期，城市管理走到数字经济时代后，可以借助的措施更多。但是在这次新冠肺炎疫情防控过程中，我们发现已经开展了10余年的智慧城市工程起到的作用非常微弱，这充分说明，我们的建设内容和投资方向并没有满足城市的基本需求。

在数字经济时代，依靠数字技术（数字化、互联网、数字孪生、人工智能等）发展路线，城市运转需要与之匹配的硬件系统、软件系统、法制与机制。我们在建设这些内容时应该以城市的生存为基础、为最大公约数，比如人口的生存、城市的存亡等，以此为最大公约数，作为基本法制制定好最底层的安全保障，然后在此基础上推进城市的治理、公共服务，最后实现人、企业、城市管理者、城市大脑四者之间充分协同，达

到城市组成成员更有尊严、更有荣耀。因此，智慧城市的顶层设计模型建设，应更多关注（如图2所示）发展轴中智慧城市建设需求层次，以智慧生存为最大公约数循序渐进开展，扎扎实实地分步骤实施。

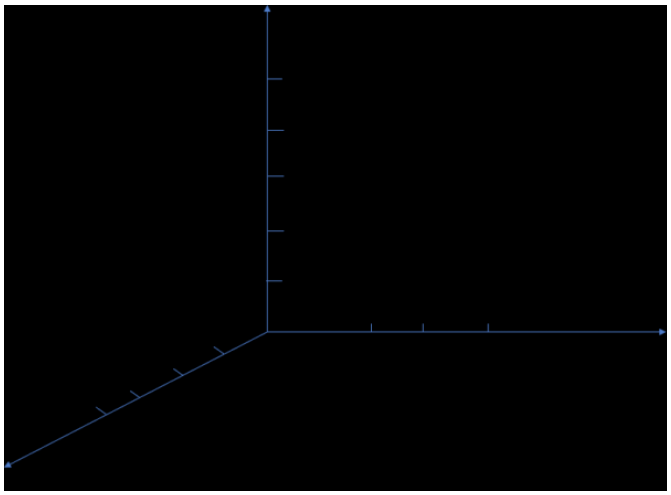


图2 智慧城市顶层设计三维模型图

**智慧城市建设的第一阶段任务：智慧生存**

生存是城市发展的前提。如果一个城市在应对各种灾害中难以生存，那么“发展”这个命题就不存在，“智慧”更无从谈起。城市存在的潜在威胁也很广泛，历史上也不乏此类案例。城市需要结合城市所处地理环境、面临的风险等级分轻重缓急去建设。例如，如果城市处在地震带上，就需要建设一套震前预报、震后救援的地震智慧应急系统，而且要保障地震后通信系统不会马上中断，并能应用于救援通信。针对大规模传染病，每个城市都面临共同的威胁，需要高度协同的疫情智慧应急系统，从而可以利用疫情智慧应急系统快速定位传染源、隔离初始感染者和潜在接触者等。

智慧城市第一阶段的建设任务，形成智慧城市的最基础应用系统——城市生存保障系统。该系统包含城市应对各种灾害时的各个子系统，比如疫情控制系统、地震预报与处置系统、大规模水体污染预警与处置系统、高浓度空气污染预警与处置系统等。

第一阶段是用最简单有效的方式实现城市生存，数字化、互联网、数字孪生、人工智能需不断升级，各种应急举措和系统维护应该以立法的形式加以固定。

**智慧城市建设的第二阶段任务：智慧治理**

城市生存下来才能谈治理。城市的土壤环境、绿化、水体、市容市貌、空气、交通、公共安全、医疗保障、购物环境等公共设施等都需要城市管理者开展治理，给市民的吃、穿、住、用、行提供硬件的便利，这属于城市规划和市政建设范畴。

智慧治理是在第一阶段城市生存任务的基础上，通过数字化不断优化城市三大环境，回顾这10余年的智慧城市建设，从各地政府对智慧城市和数字政府的建设理念和定位来看，大多是信息化项目，缺乏从提升城市治理体系和治理能力的现代化角度来认知，建设“城市大脑”“城市指挥控制中心”等，过多强调“形象展示”“高新技术”。

智慧治理分为数字化、互联网化、物联网化三大阶段。在数字化阶段，传统的人工标签、模拟设备模式要被数字化设备取代，只有统一了数字制式，才能互联互通；在互联网化阶段，智慧治理通过协同各部门间的数字化信息做好相关决策；在物联网化阶段，智慧治理通过物联网打通物理环境和城市管理活动。

**智慧城市建设的第三阶段任务：智慧服务**

智慧服务是在智慧治理的基础上为企业提供营商的全流程服务，优化企业营商环境，为企业提供财税、工商、安全、品牌打造、知识产权保护、人力资源引进等各类便捷平台；为市民办理政务提供高效、便捷的服务，无论市民“漫游”在世界哪个角落，这种智慧服务使市民具有城市归属感和认同感；对于城市管理者来说，智慧服务能够为他们提供高效协同的一体化平台，而不用再把精力消耗在协调各个厅、局、处、科室上，减少“文山会海”，城市管理者将有更多精力去策划更先进的城市管理思路，提升响应速度，提升服务水平。

比如，一个城市有很多家医院，但是每个市民一生中可能会去多家医院就诊，那么市民就需要一个电子病历，如果到任何一家医院，只要病人现场授权（密码、指纹、扫脸等）就可以让医生调阅其既往症和跨医院诊疗数据、医疗影像等，这样将大大提升城市服务水平。但目前为止，中国还没有哪个城市能做到，假如某一个城市做到了，那么就会增加市民对其的归属感。智慧服务要体现在每个具体难题的解决上。

**智慧城市建设的第四阶段任务：智慧发展**

只要城市中的企业能够融入当地产业，通过政府引导创新发展，为当地贡献更多的就业、财税，带动当地下游企业，那么就会带来持续的发展动力。从企业层面来说，智慧发展需要产业集群、工业互联网、产业互联网的支持，筑巢引凤，吸引更多的世界500强企业、民营100强企业入驻。很多服务并不一定是政府“大包大揽”，而是明确发展理念，做好引导服务，让更多的社会企业参与到智慧城市的建设中。由此，“智慧服务”会衍生出更多的社会需求，进一步拉动智慧发展。政府在这中间应该营造公平的竞争环境，并鼓励社会企业多参与。

从市民角度来看，城市从教育、就业等方面提供社会供给，吸引全球人才到城市工作、生活。这需要智慧化系统能够拉近城市与每个市民的距离，让每个人都能通过平台找到自己的发展路径。这不仅是解决最基础的就业问题，中高层人才的信息化管理也是城市可持续的关键要素。

对于城市管理者而言，数字城市-无线城市-光网城市-互

联网+城市-数字孪生城市-城市大脑，这是一条智慧城市软硬件发展的主线，管理者找到发展主线才能实现智慧发展。

**智慧城市建设的第五阶段任务：智慧荣耀**

根据世界银行的定义，当前我国一线城市已经达到了发达国家的生活水平。当人、企业、城市实现了充分自由的发展，每个人、每个企业的发展潜力与城市的发展完美融合，人的发展、企业的发展其实都实现了自我价值，并且在发展过程中带动了周边城镇、周边经济的发展，这种被模仿、被强吸引的能力就是实现了城市的荣耀。

在这个阶段，城市彻底实现了数字孪生城市、人工智能城市。由于人力资源是较稀缺资源，因此未来人们的自由发展建立在人工智能上，实现高效率、高质量、安全、节能、环保的运营。但目前全球还没有哪个城市能做到这种程度。

当前的“城市大脑”严格意义上不是人工智能，仍然是属于城市发展的第二阶段——治理城市交通拥堵的智慧交通系统。以交通为例，在城市的智慧荣耀发展阶段，交通完全基于人工智能的无人驾驶、智能网联汽车，车辆之间的路口交通不再需要红绿灯管控。

**智慧城市建设等级理论新愿景**

相比于中华上下五千年的文明史，智慧城市的建设周期虽然有10余年，但还非常短暂，智慧城市的建设也正在不断试错中摸索，但如果有演进路径就可以减少成本浪费。基于上文，笔者将智慧城市的发展分为5个级别（如表所示）。

1星级智慧城市：城市在应对自然灾害、大规模传染病等威胁场景下，能够有较大的敏捷性，使生活在城市中的人们损失最小化。比如在类似2020年新冠肺炎疫情中，如果某市突发传染病，智慧城市会及时发现、及时隔离，不会扩散蔓延到多地区甚至全国。那么如何从智慧城市的角度应对出现新疫情的情况？这就需要城市通过智慧生存系统保持城市的健壮。

2星级智慧城市：使城市在交通、大气、水体、社区治安、生活物资供应方面保持友好，实现治理的数字化、互联网化、平台化，应用于各类垂直行业的管理，城市实现了现代工业化。

3星级智慧城市：在城市治理的基础上，为城市市民、企业服务，达到政通人和，促使城市有独特的文化和服务体系，市民的归属感增强。3星级智慧城市拥有一系列的服务平台，并可以广泛利用移动互联网和物联网服务。

4星级的智慧城市：人尽其才、物尽其用，互联互通，城市走在高质量、相对中高速增长的快车道，数字经济高度发达。

5星级智慧城市：实现人工智能城市，相比于数字经济时代，工业时代的工作量基本被人工智能取代，数字经济时代的工作量有50%被人工智能取代。5星级智慧城市物质高度丰富，人们可以摆脱经济瓶颈与物质限制，实现自由自主生活。

| 建设等级 | 星 等 | 级别 | 关键的智慧城市系统工程 |
|---|---|---|---|
| 1星级智慧城市 | ★ | 生存级 | 自然灾害应急处置系统、疫情应急防控系统…… |
| 2星级智慧城市 | ★★ | 治理级 | 智慧环保、智慧医疗、智慧交通…… |
| 3星级智慧城市 | ★★★ | 服务级 | 智慧政务、智慧人社、智慧工商…… |
| 4星级智慧城市 | ★★★★ | 发展级 | 智慧教育、工业互联网、产业互联网、数字孪生城市 |
| 5星级智慧城市 | ★★★★★ | 荣耀级 | 城市大脑、人工智能普及到各产业 |

表 智慧城市发展的5个级别

总结反馈是智慧的表现。疫情是一场战役，在面临生死存亡的极端时刻，很多积压的供需矛盾、走入误区的做法常常容易爆发出来。如果这场疫情能让我们好好总结智慧城市以及众多管理中的突出问题，并且在今后加以解决并改正，相信在一定程度上也可以弥补一些因疫情带来的损失。本文也是基于此，抛砖引玉，关于五星级发展路径中的前后演化阶段中的技术体系之间的关系还需深入分析，仅供借鉴。

# / 企业如何做生产设备更新决策 /

来源 / 广西CPDA数据分析师 陈虹坚　编辑 / 协会会员处 李苗苗　日期 / 2021-02

在企业的日常运营及经营管理中，生产设备能够高效、经济地运行一直是设备管理者关注的问题。在用设备使用了多年，虽然未到报废期，但因其磨损程度提高、设备性能下降，导致的运维费用逐年增长；对于是否需要更换新设备，如何进行设备更新决策，是本文主要探讨的问题。

**一、需求背景**

某公司计划采用新设备替代旧设备，旧设备当前已使用3年，设备原价15000元，目前变现价值7500元，估计尚可使用5年，预计残值1800元，每年维修成本1500元，付现操作成本900元；新设备购置成本12500元，预计可以使用6年，预计最终残值2000元，每年维修成本1300元，付现操作成本800元。

该设备采用直线法计提折旧，折旧年限6年，残值率为10%，该公司预期报酬率10%，所得税税率25%。新旧设备数据对比如下表：

| | 原价 | 已使用年限 | 尚可使用年限 | 当前变现价值 | 预计残值 | 每年维修成本 | 付现操作成本 |
|---|---|---|---|---|---|---|---|
| 旧设备 | 15000 | 3 | 5 | 7500 | 1800 | 1500 | 900 |
| 新设备 | 12500 | 0 | 6 | 12500 | 2000 | 1300 | 800 |

**二、解决思路**

该公司需要进行设备的更新决策，决定是否需要采用新设备替代在用的旧设备。我们可以思考一下，设备的使用产生的成本分为固定资产成本和运行成本，固定资产成本的构成分别为初始投资、折旧抵税和残值变现；运行成本的构成为操作成本、维修费用和残次品成本。其中，继续使用旧设备的初始投资成本是当前进行旧设备处置的税后现金流，即初始投资=当前账面价值*税率+当前变现价值*（1-税率）。

我们目前需要决策的事项是在已知的背景条件下，应该继续使用旧设备还是更换新设备，即该公司目前有两个方案，方案一是继续沿用旧设备；方案二是使用新设备；我们只需要比较两个方案哪个更经济实惠即可。因此，只需要分别计算出继续使用旧设备和购买使用新设备的平均年成本，对两个方案的平均年成本进行比较即可。

**三、计算步骤**

1.计算旧设备资产成本

旧设备原价15000元，已经使用3年，折旧年限6年，残值率为10%。

则有旧设备当前账面价值：15000-15000*(1-10%)/6*3=8250元

已知当前变现价值：7500元

则有初始投资成本：8250*0.25+7500*0.75=7687.5元

该公司所得税税率25%

固定资产折旧抵税：15000*(1-10%)/6*25%=562.5元

旧设备已使用3年，折旧年限6年，尚可享受3年折旧抵税。

第5年固定资产残值变现：

第5年账面价值15000*10%=1500元

第5年变现价值1800元

残值变现1500*0.25+1800*0.75=1725元

2.计算旧设备运行成本

每年维修成本1500元，付现操作成本900元
每年税后付现成本(1500+900)*(1-25%)=1800元

3.计算继续使用旧设备的平均年成本
现值总成本7687.5-562.5*（P/A,10%,3）-1725*(P/F,10%,5)+1800*(P/A,10%,5）=12041.00625元
继续使用旧设备平均年成本12041.00625/(P/A,10%,5）=3176.37602881

4.计算使用新设备的资产成本
新设备购置成本12500元，折旧年限6年，残值率为10%，该公司所得税税率25%。
固定资产折旧抵税：12500*(1-10%)/6*25%=468.75元
残值变现：
账面价值12500*10%=1250元
变现价值2000元
残值变现1250*0.25+2000*0.75=1812.5元
5.计算新设备运行成本
每年维修成本1300元，付现操作成本800元。
每年税后付现成本(1300+800)*(1-25%)=1575元

6.计算使用新设备的平均年成本
现值总成本12500+(-468.75+1575)*（P/A,10%,6）-1812.5*(P/F,10%,6)=16294.894375元
使用新设备平均年成本16294.894375/（P/A,10%,6）=3741.39424954元

7.比较两个方案的平均年成本
继续使用旧设备平均年成本3176.37602881元
使用新设备平均年成本3741.39424954元

8.结论
继续使用旧设备平均年成本低于使用新设备平均年成本，因此该企业应该选择沿用旧设备的方案。

**四、小结**

企业判断一台正在使用的设备是否需要更新，做更新决策之前，应该全面地了解现有设备的性能、使用年限、维护费用等运行成本，以及新设备的技术进步等因素，结合财务管理的知识，选择更经济的方式，做出设备更新的最优决策。

# / 描述统计分析应用的感触 /

来源 / CPDA数据分析师 刘程浩　编辑 / 协会会员处 李苗苗　日期 / 2021-02

这段时间忙了一阵子数据可视化的工作，有个很深的感触：一直以来，我觉得比较入门级的描述统计，用的好还是可以帮助业务部门发现问题，甚至帮助我们的客户确定未来的发展方向的。在写正文之前，我先说下我觉得比入门级更高级别的数据分析，甚至觉得很高大上的数据分析是什么？我认为是依赖深奥的数学理论支撑的各种建模分析。举例子来说，最简单的线性回归，略微复杂的时间序列回归模型ARIMA，带有可动态调整能力的状态空间方程、神经网络模型等等。

可能也和过去的业务特点有关，因为涉及到对未来的预测，因此这些建模分析工具需要捕捉到数据中的特征，并且用一些系数、参数、超参数来对过去的历史规律进行控制和表达。

不过，这些建模工具有个特点，就是一旦模型确定后（系数、参数、超参数确定），往模型里添加什么数据，跑出来的结果就牢牢地印上了历史规律的特点。这样一来对未来的预测往往就确定了。打个形象的比喻，模型就好比模具。像线性回归模型，无论是做何种线性变换进行数学公式的改变，经过历史数据的计算，一旦系数定了，那么模型想要表述的规律就定了，一点儿也不能改。

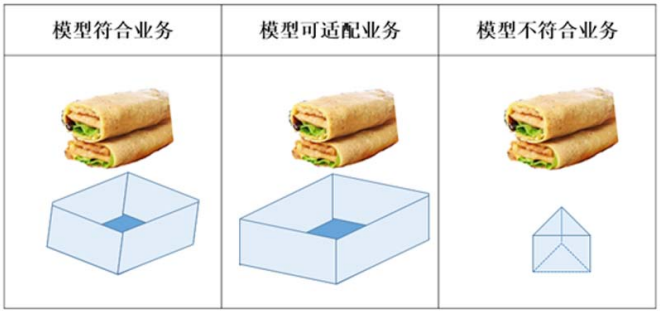
例如下面的月饼模具，因为是用木头雕刻的，所以一旦雕刻好，那么半径定了，花纹定了，无论你用啥糯米和颜料进行装饰，打出来的月饼就是那个样子。

而那种带有自动调节功能的模型，例如卡尔曼滤波用到的状态空间方程、各种神经网络模型，他们都有一些柔性，面对最后一批次的样本数据会调整自己的系数。这样一来模型的预测就会根据最后一段时间的数据规律或数据特征进行调整。

打比方的话，就像高分子模具例如硅胶磨具、或者再生活化一点诸如装水的气球。因为如果多装一点或少装一点原材料，模具会根据材料的多少发生形变，这样能够生产出形状略有差异的产品。

以上我说的建模的数据分析模型，背后都有比较深奥的数学知识在做支撑。而想要掌握和应用自如，还离不开各种分析软件、工具。自己还要学会对参数的调整、搞明白背后的数学逻辑后，懂得对模型进行选择和适配。因此对数学功底的要求比较高。但这些模型在使用时，和业务部门进行解释往往比较吃力。比方说，什么叫系数矩阵的特征值，光特征值这个词要想和日常生活里的“特征”进行区别解释，也是不容易的了。

除此之外，模型拟合出来的效果，乃至于预测出来的预测值，和实际业务进行对比起来，也有差异。就如同下面的3个图中的盒子，用它们来比喻我们的模型；图中的煎饼，就好比我们的业务。我们追求的是最左边的效果；然而实际上能做到中间的效果就是能让自己和业务部门都能接受的了；但是往往在建模分析的早期，效果是最右边图的呈现那样。




俗话说没有对比就没有收获，以上的感触和心得，是在我最近不断的使用数据可视化工具对数据进行描述统计分析的时候得到的。描述统计分析一般就是分析均值、方差、分布等。一般人看方差还看不出个所以然，所以很多人更喜欢看容易理解的均值。

如果单独看均值，说实话没有经过对比也是看不出什么结果的。但是如果拿到可对比的参考物前进行对比，那就有意义的多了。

比方说，我在网上搜到了2015年日本厚生省统计的学龄前儿童的身高体征的均值，拿去我家小孩所在的幼儿园家长群里，或者拿到我的亲戚、同事群里，好多人通过对比都说自己的孩子身高和体重超过了日本同龄人。

Japanese Boy

| 5 year-old | cm | feet | SD | kg | pounds | SD |
|---|---|---|---|---|---|---|
| 0 months | 106.7 | 3' 6" | 4.3 | 17.7 | 39.0 | 2.5 |
| 1 months | 107.3 | 3' 6" | 4.4 | 17.9 | 39.5 | 2.6 |
| 2 months | 107.8 | 3' 6" | 4.4 | 18.1 | 39.9 | 2.8 |
| 3 months | 108.3 | 3' 7" | 4.4 | 18.3 | 40.3 | 2.9 |
| 4 months | 108.9 | 3' 7" | 4.5 | 18.5 | 40.8 | 2.9 |
| 5 months | 109.4 | 3' 7" | 4.5 | 18.7 | 41.2 | 2.9 |
| 6 months | 110.0 | 3' 7" | 4.5 | 18.9 | 41.7 | 3 |
| 7 months | 110.5 | 3' 8" | 4.6 | 19.1 | 42.1 | 3 |
| 8 months | 111.1 | 3' 8" | 4.6 | 19.3 | 42.5 | 3 |
| 9 months | 111.6 | 3' 8" | 4.7 | 19.6 | 43.2 | 3 |
| 10 months | 112.2 | 3' 8" | 4.7 | 19.8 | 43.7 | 3.1 |
| 11 months | 112.7 | 3' 8" | 4.7 | 20.1 | 44.3 | |

Japanese Girl

| 5 year-old | cm | feet | SD | kg | pounds | SD |
|---|---|---|---|---|---|---|
| 0 months | 106.2 | 3' 6" | 4.2 | 17.4 | 38.4 | 2.3 |
| 1 months | 106.7 | 3' 6" | 4.3 | 17.6 | 38.8 | 2.4 |
| 2 months | 107.3 | 3' 6" | 4.3 | 17.8 | 39.2 | 2.5 |
| 3 months | 107.8 | 3' 6" | 4.3 | 18 | 39.7 | 2.6 |
| 4 months | 108.4 | 3' 7" | 4.4 | 18.1 | 39.9 | 2.6 |
| 5 months | 108.9 | 3' 7" | 4.4 | 18.2 | 40.1 | 2.6 |
| 6 months | 109.5 | 3' 7" | 4.4 | 18.4 | 40.6 | 2.7 |
| 7 months | 110 | 3' 7" | 4.5 | 18.5 | 40.8 | 2.7 |
| 8 months | 110.6 | 3' 8" | 4.5 | 18.6 | 41.0 | 2.7 |
| 9 months | 111.1 | 3' 8" | 4.5 | 18.7 | 41.2 | 2.8 |
| 10 months | 111.6 | 3' 8" | 4.6 | 19 | 41.9 | 2.8 |
| 11 months | 112.2 | 3' 8" | 4.6 | 19.3 | 42.5 | |

还有，例如2020年的疫情期间，通过国际第三方数据公司获取到的某些区域的网速热力图，放到一个时间轴上，就可

以看出区域间的网速变化。

结合当时国内外相互借鉴的封城、居家办公、远程会议……等等措施，我们就能够很轻松得到一个结论：疫情改变了我们的网络使用强度和深度，并且随着疫情的控制，这种网络使用模式并没有回退到过去最初的状态，并且需要国家和社会持续加大互联网软硬件的建设投入。这个结论拿到客户面前看，客户心里面对未来的预测，就相对容易多了。

下面是我用比较粗糙的方法模拟出来的区域城市网络下载平均速率的热力图，从热力图中我们很容易向读者展示以下的信息：

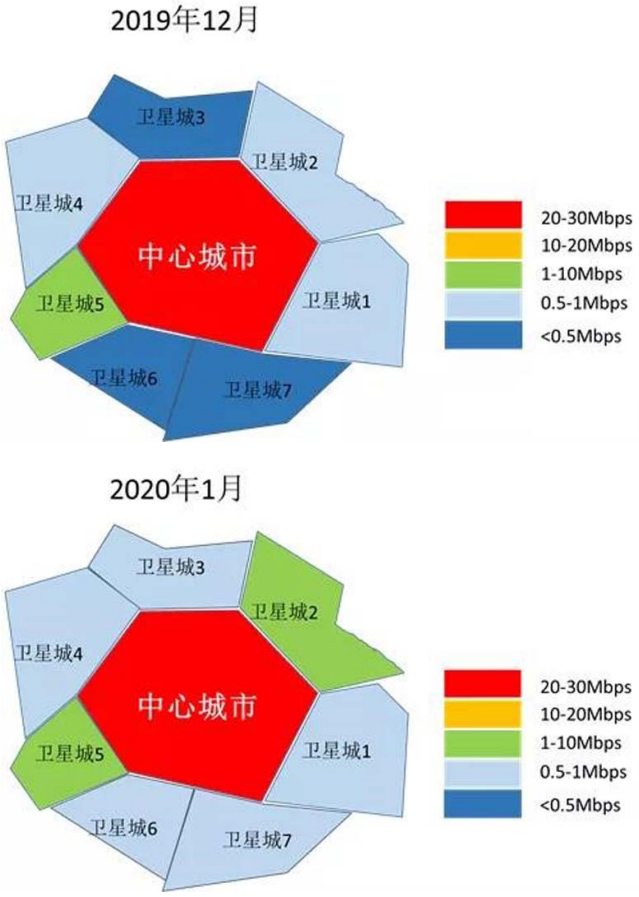
1. 中心城市，由于商业和人口都比较密集，因此大家对网络的使用强度和深度是刚性的，无论疫情变化如何。

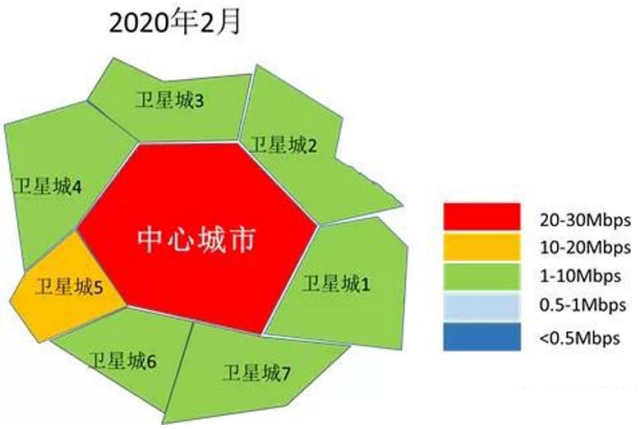
2. 对于卫星城市，由于封城和居家办公之前，商业和人口都没有那么密集，因此不同的卫星城市显然网络使用的强度和深度都是远不如中心城市的。

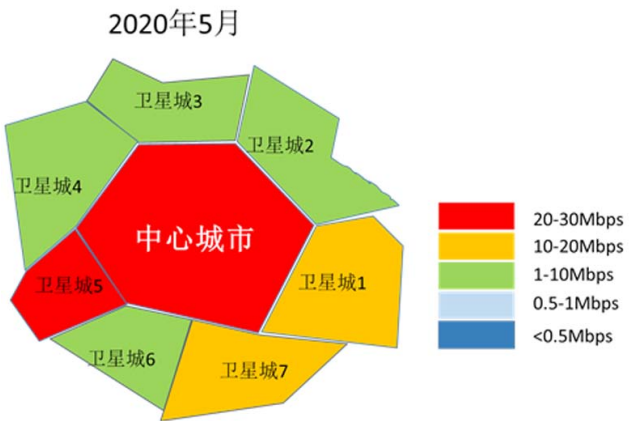
3. 由于疫情的发展导致了封城和居家办公，人们的网络使用方式发生了重大改变，卫星城市的人们对网络使用的强度和深度都逐渐接近中心城市。

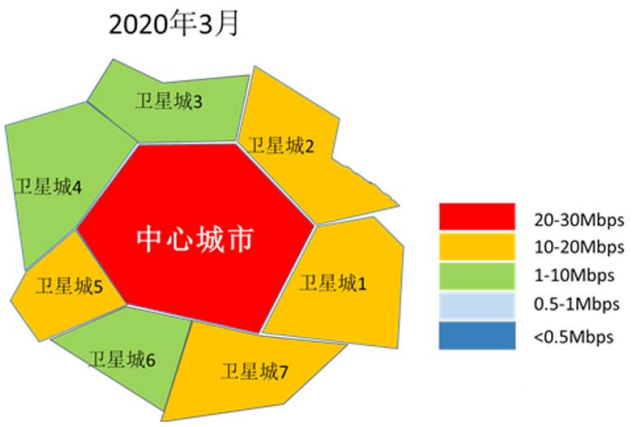
4. 随着疫情被控制，虽然人们在中心城市和卫星城市之间的流动恢复了，但是人们在商业上的网络使用强度和深度却没有马上变回疫情之前的状态，反而缓慢的变化，并且在距离疫情初期较大距离处稳定下来。

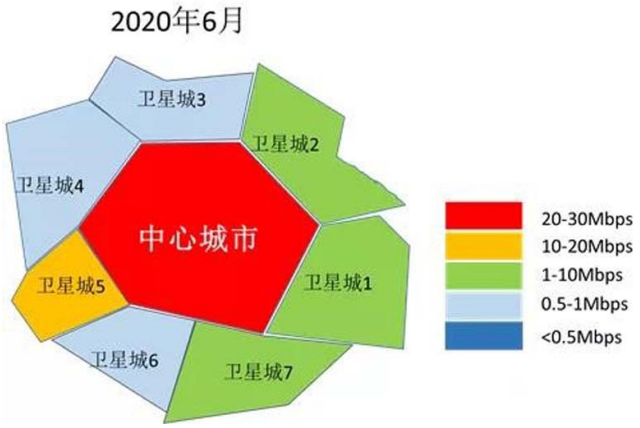
5. 因此，作为远程办公的设备商、网络服务提供的运营商，都不能认为疫情的控制代表了网络及其周边市场需求的昙花一现，而是要更加的提升自己的产品和服务的质量（用户体验），在新的对手出现之前进一步去争抢市场份额。

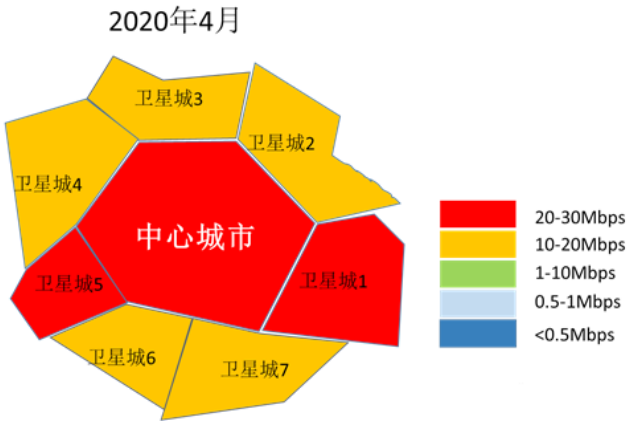





以上的数据可视化方法比较粗糙，如果大家能用支持热力图或区块图的软件或者工具，在一个固定的图上拖动时间轴看色块的变化，是很容易得到我刚才的结论的。当然了，还有其他形式的描述统计分析我就不再在这里呈现了。我觉得伟大的前人有句话说的好：问题表述清楚了，就解决了一半。描述统计分析，就能轻松和直观地起到这个效果。

# / 基于大数据的用户流失及预警模型的构建研究 /

来源 /CPDA数据分析师 马苏楠 编辑 / 协会会员处 李苗苗 日期 / 2021-02

我是一名中国联通武汉分公司基层员工，工作中涉及大量业务数据，一直对数据分析在企业营销中的应用十分感兴趣，但缺乏数据分析的专业技能，2020年8月我报名参加了武汉第26期CPDA培训认证，获得了CPDA证书。

通过对CPDA课程的学习，让我对大数据有了充分认识，建立了大数据分析思维，熟悉了数据分析流程和方法，掌握了建模技能和算法，从普通的数理统计提升为用专业数据分析技能挖掘数据价值，应用和指导实际业务。特别是课程中将整个营销过程所包含的数据分析理论和方法全部涵盖了进去。

其中客户画像就是通过对消费者基本信息的搜集，提取关键信息，同时利用逻辑回归，判断数据之间的相关关系，以关系来推断消费行为与消费走向，从而做到精准营销。数据分析的培训课程在我的工作中得到了很好的实际运用。

以下内容是笔者结合工作实际，利用CPDA课程知识与DATAHOOP智能分析平台，编写的本篇论文，在一定程度上可以拓展营销领域的分析视角。

摘 要：本文利用大数据平台构建了用户流失及预警模型，通过采集移动运营商的用户行为数据，并对数据样本进行预处理，利用机器学习等技术方式，计算出用户流失及预警的逻辑回归模型，以期可以通过该模型，让运营商政策制定者提前制定各项维系政策，达到精准化、精细化运营。

关键词：大数据；机器学习；流失预警；精准化运营

引言：2019年6月，我国正式进入5G商用元年。5G的高链接速率，大大提升了人们对“网络”的感知，同时，通过联网，人类也在不断主动的创造数据，直到各行各业在应用系统上的规模不断扩大，数据也呈爆炸性增长，人类已正式进入大数据时代。

随着计算机和信息技术的迅猛发展和普及应用，QQ、微信、支付宝、淘宝、抖音等以流量消耗为主的APP逐渐占据人们的生产生活，通信运营商与各类互联网技术公司业务开始深度融合，使得移动通信技术有了纵深的发展，也不断改变着用户的通信、交流、沟通方式。

各类业务使用的软件类型、流量类型、终端类型，甚至到移动用户发出的每一条指令，都逐步累积为海量的用户数据。与此同时，各类通信运营商为了可以更好的利用大数据的工具，对用户数据进行充分挖掘，在横向纵向上充分完整的分析用户行为，营造了良好的维系环境。

经过几十年的发展沉淀，三大运营商在融合与竞争中，用户逐步趋于饱和，拓新的广度与难度也在逐渐扩大。与此同时，国务院提出“提速降费”、“光纤入户”等政策，使得运营商的建设成本和营销费用不断上升，使用高新技术必然导致运营成本提高，在这种情况下，运营商的竞争，不只局限于拓新用户上，自身存量用户市场的保有份额也成为了竞争的另一个重点。

根据前期市场已有统计数据可以看出，三大运营商在拓新用户上花费的平均成本为300元左右，拓新后，用户号

码的使用保有率不足30%，而维系一个在网老用户花费的成本却不到45元，在老用户选择维系产品后最终留下来不退网的生命周期远大于一个新用户的发展。因此，关注老用户维系，拓展老用户生命周期也逐渐变成目前各大运营商需要解决的焦点工作。

当下，数字营销已经深入人心，运用数据进行指导并提升营销效果也变成了营销的重要手段。数据作为各种统计、计算、科学研究或技术设计等所依据的数值，在日新月异的信息时代，其作用日益凸显，作为移动通信运营商，用户的各项终端、流量、通话、详单数据已不仅仅是数值的呈现，更是各大通信运营的核心资产。

而以往处理用户数据的方式是通过各类数据库软件或者编程进行分析，再制定针对不同用户的营销策略，这种方式往往无法充分进行用户细分，软件编程的准确性和及时性都有很大的差异。5G技术也将更多的营销场景和营销渠道进行整合，可以更精准地了解到用户的行为，实现精准维系。

本文利用大数据基础平台，对通信运营商的用户流失情况进行研究，建立用户流失预警系统，通过充分利用数据挖掘和机器学习方法，对各类用户数据进行学习研究，力求准确科学的建立用户流失预警模型，为差异化营销及用户挽留提供精准的数据支撑。

用户流失预警系统可分为用户基础信息与属性信息、数据预处理、构建模型并调整优化、模型评估和输出结果。

用户流失一般是指用户已停止使用某个通信运营商的某项产品长达数月，转而选择使用其他通信运营商的产品，本文将用以下概念定义流失用户：

用户长时间不使用产品，账户内出现欠费长达3个月，达到即将销户的状态，或者用户账户标识为销户，则定义为用户流失。

## 一、数据选取

由于日常接触到的用户数据类型较多，研究用户流失预警系统就需要有针对性的选出与用户流失有强相关的且显著性较强的字段与变量，从而确定模型。按照用户的基础属性信息、消费使用行为信息等数据，选出用作研究流失特征的指标变量。

| 数据字段属性分类 | |
|---|---|
| 基本信息 | 年龄、性别、在网状态、在网时长、套餐类别、参与优惠 |
| 消费信息 | 月出账、通话费用、流量费用、增值业务费、短信条数、流量使用 |
| 使用信息 | 当月通话次数、当月通话时长、主叫通话时长、被叫通话时长、国际漫游次数 |
| 缴费信息 | 首次缴费金额、累计缴费金额 |

## 二、数据预处理

本次研究选取了5000名用户在1个月内的通话、流量、缴费等数据。这些数据来自多个数据库，有些数据名称不一致、存在异常值，无法直接在平台中建模，需要先将数据进行归类和预处理，对原始数据进行清洗。

（1）异常值处理。

三无（无语音、无流量、无短信）和极低（通话时长在10 分钟以内和流量消耗在10M 以内）的用户剔除。

（2）特殊值处理。

员工套餐及宽带、政企用户因享有群体特殊优惠，不能作为单个用户对待，对流失预测特征的研究意义不大，故删去。

（3）数据变换。

由于一些变量指标数据平台无法直接读取，需要把字符型数据数值化，所研究变量中的流失状态、性别需要做相应转换。

打标系统状态为“销户”的为“1”，未销户的其他状态打标为“0”。

用户性别标注男性为“0”，女性为“1”。

经过以上异常值、特殊值处理和数据变换后，最终有3500名用户的数据适用于本次建模研究。

（4）将确定下来的指标变量进行定义

| 研究选取指标变量定义列表 | | | | |
|---|---|---|---|---|
| 变量 | 变量类型 | 变量名称 | 变量标签 | 定义 |
| 因变量 | 分类 | IS_L | 是否流失 | 0：未流失 1：流失 |
| 自变量 | 连续 | X1 | 年龄 | |
| 自变量 | 分类 | X2 | 性别 | 0：男性 1：女性 |
| 自变量 | 分类 | X3 | 在网状态 | 0：正常 1：欠费 |
| 自变量 | 连续 | X4 | 在网时长 | 已入网月份数 |
| 自变量 | 分类 | X5 | 套餐类别 | 用户套餐分类为： A1：1，B1：2，C1：3 D1：4，E1：5，F1：6 |
| 自变量 | 分类 | X6 | 参与优惠 | 0：有 1：无 |
| 自变量 | 连续 | X7 | 月出账 | 当月出账费用（单位：元） |
| 自变量 | 连续 | X8 | 通话费用 | 当月语音使用费用 （单位：元） |
| 自变量 | 连续 | X9 | 流量费用 | 当月 GPRS 使用费用（单位：元） |
| 自变量 | 连续 | X10 | 增值业务费 | 当月增值业务费用 （单位：元） |
| 自变量 | 连续 | X11 | 短信总数 | 当月收发短信条数 |
| 自变量 | 连续 | X13 | 流量使用 | 当月 GPRS 使用量（单位：MB） |
| 自变量 | 连续 | X14 | 主叫通话时长 | 当月主动拨打出去的时长 （单位：分钟） |
| 自变量 | 连续 | X15 | 被叫通话时长 | 当月接收电话的时长 （单位：分钟） |
| 自变量 | 连续 | X16 | 国际漫游次数 | 当月使用国际漫游通话的频率 |
| 自变量 | 连续 | X18 | 首次缴费金额 | 入网当月缴费金额 |
| 自变量 | 连续 | X19 | 累计缴费金额 | 截止目前用户累计已缴费金额 |

定义完相关变量后，按照数据分类方法，把3500名用户按照6:2:2的比例将数据分为训练集、测试集和验证集。其中，训练集有2125名用户信息，测试集有721名用户信息，验证集有654名用户信息，符合数据建模标准。

## 三、构建模型并调整优化

因为本次研究所选取的变量指标较多，无法准确的判断哪一个自变量类型与因变量有较强的相关性。因此需要在自变量与因变量的相关性上进行分析和验证，从而将相关性较弱的自变量删除，达到最优模型结果。

采用卡方检验验证，将训练集数据导入数据分析平台，选择卡方检验，逐步筛选出显著性值小于0.05的自变量，优化数

据结构。分别对选取的自变量数据进行检验，得出以下结果：

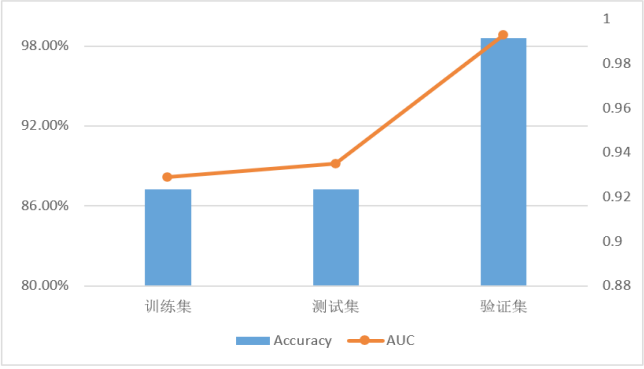
| 变量 | 名称 | 卡方值 | Sig 值 |
|---|---|---|---|
| X1 | 年龄 | 4.25 | 0.073 |
| X2 | 性别 | 2.348 | 0.365 |
| X3 | 在网状态 | 2.1826 | 0.282 |
| X4 | 在网时长 | 5.933 | 0.041 |
| X5 | 套餐类别 | 4.152 | 0.642 |
| X6 | 参与优惠 | 1.984 | 0.021 |
| X7 | 月出账 | 4.125 | 0.032 |
| X8 | 通话费用 | 5.79 | 0.374 |
| X9 | 流量费用 | 5.452 | 0.162 |
| X10 | 增值业务费 | 8.08 | 0.573 |
| X11 | 短信总数 | 23.51 | 0.382 |
| X13 | 流量使用 | 5.987 | 0.012 |
| X14 | 主叫通话时长 | 4.857 | 0.013 |
| X15 | 被叫通话时长 | 6.772 | 0.036 |
| X16 | 国际漫游次数 | 7.982 | 0.375 |
| X18 | 首次缴费金额 | 4.517 | 0.829 |
| X19 | 累计缴费金额 | 5.983 | 0.034 |

由上表得出，自变量为X4、X6、X7、X13、X14、X15、X19显著性值小于0.05，说明这7个自变量与因变量“是否流失”IS_L间存在较强的相关性，显著性较强，可以作为回归模型的变量，参与模型运算。其他自变量指标优于大于0.05，说明与因变量间没有显著的相关性，对因变量的影响较小，因此不带入模型进行运算。

选取训练集中自变量为在网时长、参与优惠、月出账、流量使用、主叫通话时长、被叫通话时长、累计缴费金额共7个自变量进行Logistic 回归分析，求出用户流失预测的模型方程为：

E（y）=65+21.92*在网时长＋10.09*参与优惠＋11.36*月出账-9.34*流量使用+10.57*主叫通话时长+9.73*被叫通话时长+10.79*累计缴费金额

模型在流失用户的预测准确率上表现为87.25%，AUC即ROC曲线下与坐标轴围成的面积值为0.929。




同样，使用相关自变量对之前分类出来的测试集进行模型演算检验，算出模型在流失用户的预测准确率为87.27%，AUC即ROC曲线下与坐标轴围成的面积值为0.935；使用验证集对模型方程进行验证，得出流失用户的预测准确率为98.6%，AUC即ROC曲线下与坐标轴围成的面积值为0.993。

通过对数据的训练、测试及验证，得出各项准确率指标，表明，此次建模方案可行，对移动运营商的存量用户维系工作能够带来一定的参考，后台维系方案政策制定者可以根据与用户是否流失有较强显著性指标的变量对用户进行针对性的挽留，提高维系效果。

## 四、精准化运营

利用回归模型及其相关自变量参数与回归方程，可以总结出本次研究有如下结果：

（1）用户在网时长越长，流失的概率越小；

（2）用户参与了相关优惠，如存费送、电子券等活动，流失的概率小；

（3）用户月出账费用越高，流失的概率越小；

（4）用户在GPRS等流量上的使用越多，流失的概率越小；

（5）用户拨打和使用的通话分钟数越多越频繁，流失的概率越小；

（6）用户从开始使用到截止统计时点，缴费的金额越高，越不容易流失。

因此，移动运营商可以针对以上用户行为特点，制定以不同用户使用特点为主要特征的维系策略，来满足不同用户对通信需求，已达到用户的精准化维系。同样，移动运营商的政策制定部门，可以利用以上模型，来判断现有存量用户的消费特征，建立用户使用与流失预警模型，对异常用户进行预判。

## 五、结语

随着互联网技术的不断发展，互联互通的零距离更加成为人们的迫切需求，后疫情时代的到来，人们需求的增加也对移动通信行业有了更高水平的要求。大数据、云计算、物联网等数据挖掘和运用方式也越来越被人们熟知，充分了解用户的各种使用习惯，通过用户使用数据掌握用户的及时需求，从而有针对性的解决在运营过程中出现的各种问题，才能达到用户的精准化运营，提升5G时代的坐标站位。

# / 实战案例 | 泰坦尼克号乘客生存预测 /

作者 /数据运营与数据分析 Jalenxr 编辑 / 协会会员处 李苗苗 日期 / 2021-02

**1、什么是集成学习方法**

集成学习通过建立几个板型组合的来解决单一预测问题，它的工作原理是生成多个分类器/模型，各独立地学习和作出预测。这些预测最后结合成组合预测，因此优于任何一个单分类的做出预测。决策树过度拟合可以用剪枝或者集成学习方法的随机森林实现。

**2、什么是随机森林**

在机器学习中，随机森林是一个包含多个决策树的分类器，并且其输出的类别是由多个决策树输出的类别的众数而定。例如，如果你训练了5个树，其中有4个树的结果是True，1个树的结果是False，那么最终投票结果就是True。

· 随机：特值随机，训练集随机
· 森林：包含多个决策树的分类器

**3、随机森林的原理过程**

随机：特值随机，训练集随机。随机森林算法根据下列算法而建造每棵树：

· 用N来表示训练用例（样本）的个数，M表示特征数目。

o 1、一次随机选出一个样本，重复N次。(随机有放回的抽取，有可能出现重复的样本)

o 2、随机去选出m个特征，m << M，建立决策制，每棵树有m个特征。

· 采取bootstrap抽样(随机有放回的抽样)

**4、为何采取bootstrap抽样**

为什么要随机抽样训练?

· 如果不进行随机抽样，每棵树的训练集都一样，那么最终训练出的树分类结果也一样

为什么要有放回地抽样?

· 如果不是有放回的抽样，那么每棵树的训练样本都是不同的，都是没有交集的，也就是说每棵树训练出来都是有很大的差异的；而随机森林最后分类取决于多棵树（弱分类器》的投票表决。

**5、python实现随机森林的接口**

语法：sklearn.ensemble. RandomForestClassifier (n_estimators=10,criterion='gini',max_depth=None,bootstrap=True,min_samples_split=2,random_state=None,n_jobs=-1)

参数：

· n_estimators:integer,optional(default=10)森林里的树木数量10,50,100,...

· criteria:string,可选(default="gini")分割特征的测量方法"entropy"、"gini"

· max_depth:integer或None,可选(默认=无)树的最大

深度5,8,15,...

- max_features="auto”,每个决策树的最大特征数量
  - If"auto",then sqrt(n_features)
  - If"sqrt",then sqrt(n_features)(same as"auto")
  - If"log2",then log2(n_features)
  - If None,then n_features
  - If "float",then float*n_features
- bootstrap:boolean,optional（default=True）是否在构建树时使用放回抽样
- min_samples_split：节点划分最少样本数
- min_samples_leaf：叶子节点的最小样本数
- random_state:int/None,有必要控制随机的状态,才能重复的展现相同的结果
- n_jobs=-1:可以调用所有内核
- 其中超参数有：n_estimator,max_depth,min_samples_split,min_samples_leaf

**6、应用总结**

- 在当前所有分类算法中，具有极好的准确率
- 能够有效地运行在大数据集上，处理具有高维特征的输入样本，而且不需要降维
- 能够评估各个特征在分类问题上的重要性

**7、案例：随机森林对泰坦尼克号乘客的生存进行预测**

```
import pandas as pd
'''1 获取数据'''
path = "http://biostat.mc.vanderbilt.edu/wiki/pub/Main/DataSets/titanic.txt"
titanic = pd.read_csv(path)
display(titanic.tail(3))
```

| | row.names | pclass | survived | name | age | embarked | home.dest | room | ticket | boat | sex |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **1310** | 1311 | 3rd | 0 | Zenn, Mr Philip | NaN | NaN | NaN | NaN | NaN | NaN | male |
| **1311** | 1312 | 3rd | 0 | Zievens, Rene | NaN | NaN | NaN | NaN | NaN | NaN | female |
| **1312** | 1313 | 3rd | 0 | Zimmerman, Leo | NaN | NaN | NaN | NaN | NaN | NaN | male |

```
# 解释数据
#字段：row.names,pclass,survived,name,age,embarked,home.dest,room,icket,boat,sex
#解释：人员编号,人员等级划分,是否幸存,名字,年龄,上传地点,家庭地址,所在船舱,船票,boat（我也不知道怎么解释）,性别

# 筛选特征值和目标值
x = titanic[["pclass", "age", "sex","embarked","home.dest","room"]]
y = titanic["survived"]

'''2 数据预处理'''
x.info() #发现缺失值，
#root变量缺失值太多，不具备解释性，删除。
x = x.drop("room", inplace=False, axis=1) #删除room列

#age变量缺失值，用均值代替
x["age"] = x["age"].fillna(x["age"].mean(), inplace=False)
#embarked 、home.dest 缺失，使用上一个人的向下填充
x = x.fillna(method="ffill",inplace=False)
#home.dest 缺失值用向上填充
display(x.head(3))
```

| | row.names | pclass | survived | name | age | embarked | home.dest | room | ticket | boat | sex |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **1310** | 1311 | 3rd | 0 | Zenn, Mr Philip | NaN | NaN | NaN | NaN | NaN | NaN | male |
| **1311** | 1312 | 3rd | 0 | Zievens, Rene | NaN | NaN | NaN | NaN | NaN | NaN | female |
| **1312** | 1313 | 3rd | 0 | Zimmerman, Leo | NaN | NaN | NaN | NaN | NaN | NaN | male |

```
'''3 特征工程'''
'''需要将x的字符串转换为数值型'''
# 1) 转换成字典（可有可无）
x = x.to_dict(orient="records")
# print(x)

# 2）、数据集划分
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(x, y, random_state=11)
#display(x_train)

# 3）、字典特征抽取
from sklearn.feature_extraction import DictVectorizer

transfer = DictVectorizer(sparse=False)
x_train = transfer.fit_transform(x_train)
x_test = transfer.transform(x_test)
# print(transfer.get_feature_names()) #返回类别名称
display(x_train)
```

```
array([[47.        ,  0.        ,  0.        , ...,  0.        ,
         0.        ,  1.        ],
       [31.19418104,  0.        ,  0.        , ...,  1.        ,
         0.        ,  1.        ],
       [33.        ,  0.        ,  0.        , ...,  0.        ,
         0.        ,  1.        ],
       ...,
       [31.19418104,  0.        ,  0.        , ...,  1.        ,
         1.        ,  0.        ],
       [31.19418104,  0.        ,  0.        , ...,  1.        ,
         1.        ,  0.        ],
       [31.19418104,  0.        ,  0.        , ...,  1.        ,
         0.        ,  1.        ]])
```

'''4 模型预估器'''

```
from sklearn.tree import DecisionTreeClassifier,export_graphviz
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import GridSearchCV
import warnings
warnings.filterwarnings("ignore") #用于排除警告
estimator = RandomForestClassifier(n_jobs=-1,random_state=1)
# 加入网格搜索与交叉验证
# 参数准备
param_dict = {"n_estimators": [100,120,200,300,500,800],
              "max_depth":[3,5,8,10,15],
              "max_features":["auto","log2",None]}
estimator = GridSearchCV(estimator, param_grid=param_dict, cv=3)
estimator.fit(x_train, y_train)
```

```
GridSearchCV(cv=3, estimator=RandomForestClassifier(n_jobs=-1, random_state=1),
             param_grid={'max_depth': [3, 5, 8, 10, 15],
                         'max_features': ['auto', 'log2', None],
                         'n_estimators': [100, 120, 200, 300, 500, 800]})
```

'''5 叉验证网格搜索结果评估'''

```
#1）交叉验证网格搜索结果评估
# 最佳参数：best_params_
print("最佳参数：\n", estimator.best_params_)
# 最佳结果：best_score_
print("最佳结果：\n", estimator.best_score_)
# 最佳估计器：best_estimator_
print("最佳估计器:\n", estimator.best_estimator_)
# 交叉验证结果：cv_results_
#print("交叉验证结果:\n", estimator.cv_results_)
transfer.get_feature_names()
```

```
最佳参数：
 {'max_depth': 3, 'max_features': None, 'n_estimators': 100}
最佳结果：
 0.8272357723577235
最佳估计器：
 RandomForestClassifier(max_depth=3, max_features=None, n_jobs=-1,
                        random_state=1)
```

'''6.模型评估'''

```
# 1）直接比对真实值和预测值
y_predict = estimator.predict(x_test)
print("y_predict:\n", y_predict)
#print("直接比对真实值和预测值:\n", y_test == y_predict)

# 2）准确率、精确率、召回率、F1-score
score = estimator.score(x_test, y_test)
print("accuracy准确率为：", score)
# 查看精确率、召回率、F1-score
from sklearn.metrics import classification_report
report = classification_report(y_test, y_predict,
                               labels=[1, 0],
                               target_names=["幸存", "死亡"])
print(report)

# 3）ROC曲线和AUC值(二分类评估)
from sklearn.metrics import roc_auc_score
AUC = roc_auc_score(y_test, y_predict)  #计算AUC指
```

标：越接近1越好

print("AUC-模型稳定性:",AUC)

```
y_predict:
 [0 0 1 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 1 0 1 1 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 1 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 1 0 1 0 0 0 0
 1 1 0 1 0 0 0 0 0 1 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 0 1 0 0 0 0 0 1 1 1 0 0
 0 1 0 0 0 1 1 0 1 0 0 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 1 0 1 1 0 0 1 0 0 0 0
 0 0 0 1 0 0 1 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 1 0 1 0 0 0 0 0 1 1 1 0 0 0 1 0 1 1 0 0 1 1 0 1 0 1 0 0 0 0 0 0 0 0 0
 0 0 1 0 0 0 0 0 0 1 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 1 1
 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0]
accuracy准确率为： 0.8145896656534954
              precision    recall  f1-score   support

          幸存       0.94      0.54      0.68       123
          死亡       0.78      0.98      0.87       206

    accuracy                           0.81       329
   macro avg       0.86      0.76      0.78       329
weighted avg       0.84      0.81      0.80       329

AUC-模型稳定性： 0.7585839450627516
```

'''7.模型的保存'''

import joblib

joblib.dump(estimator,'RandomForestClassifier_titan.pkl')#模型保存

'''8再次加载使用模型'''

import joblib

from sklearn.ensemble import RandomForestClassifier

from sklearn.model_selection import GridSearchCV

estimator=joblib.load('RandomForestClassifier_titan.pkl') #模型加载

y_predict = estimator.predict(x_test) #计算预测值x_test

print("y_predict(预测值):\n", y_predict)

```
y_predict(预测值):
 [0 0 1 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 1 0 1 1 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 1 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 1 0 1 0 0 0 0
 1 1 0 1 0 0 0 0 0 1 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 0 1 0 0 0 0 0 1 1 1 0 0
 0 1 0 0 0 1 1 0 1 0 0 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 1 0 1 1 0 0 1 0 0 0 0
 0 0 0 1 0 0 1 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 1 0 1 0 0 0 0 0 1 1 1 0 0 0 1 0 1 1 0 0 1 1 0 1 0 1 0 0 0 0 0 0 0 0 0
 0 0 1 0 0 0 0 0 0 1 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 1 1
 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0]
```

# / 64个数据分析常用语 /

编辑 / 协会会员处 李苗苗 日期 / 2021-02

**本篇文章，我们来讲讲数据分析常用语**

**1、绝对数和相对数**

绝对数：是反应客观现象总体在一定时间、一定地点下的总规模、总水平的综合性指标，也是数据分析中常用的指标。比如年GDP，总人口等等。

相对数：是指两个有联系的指标计算而得出的数值，它是反应客观现象之间的数量联系紧密程度的综合指标。相对数一般以倍数、百分数等表示。相对数的计算公式：

相对数=比较值（比数）/基础值（基数）

**2、百分比和百分点**

百分比：是相对数中的一种，它表示一个数是另一个数的百分之几，也称为百分率或百分数。百分比的分母是100，也就是用1%作为度量单位，因此便于比较。

百分点：是指不同时期以百分数的形式表示的相对指标的变动幅度，1%等于1个百分点。

**3、频数和频率**

频数：一个数据在整体中出现的次数。

频率：某一事件发生的次数与总的事件数之比。频率通常用比例或百分数表示。

**4、比例与比率**

比例：是指在总体中各数据占总体的比重，通常反映总体的构成和比例，即部分与整体之间的关系。

比率：是样本(或总体)中各不同类别数据之间的比值，由于比率不是部分与整体之间的对比关系，因而比值可能大于1。

**5、倍数和番数**

倍数：用一个数据除以另一个数据获得，倍数一般用来表示上升、增长幅度，一般不表示减少幅度。

番数：指原来数量的2的n次方。

**6、同比和环比**

同比：指的是与历史同时期的数据相比较而获得的比值，反应事物发展的相对性。

环比：指与上一个统计时期的值进行对比获得的值，主要反映事物的逐期发展的情况。

7、变量

变量来源于数学，是计算机语言中能储存计算结果或能表示值抽象概念。变量可以通过变量名访问。

8、连续变量

在统计学中，变量按变量值是否连续可分为连续变量与离散变量两种。在一定区间内可以任意取值的变量叫连续变量，其数值是连续不断的，相邻两个数值可作无限分割，即可取无限个数值。如:年龄、体重等变量。

9、离散变量

离散变量的各变量值之间都是以整数断开的，如人数、工厂数、机器台数等，都只能按整数计算。离散变量的数值只能用计数的方法取得。

10、定性变量

又名分类变量：观测的个体只能归属于几种互不相容类别中的一种时，一般是用非数字来表达其类别，这样的观测数据称为定性变量。可以理解成可以分类别的变量，如学历、性别、婚否等。

11、均值

即平均值，平均数是表示一组数据集中趋势的量数，是指在一组数据中所有数据之和再除以这组数据的个数。

12、中位数

对于有限的数集，可以通过把所有观察值高低排序后找出正中间的一个作为中位数。如果观察值有偶数个，通常取最中间的两个数值的平均数作为中位数。

13、缺失值

它指的是现有数据集中某个或某些属性的值是不完全的。

14、异常值

指一组测定值中与平均值的偏差超过两倍标准差的测定值，与平均值的偏差超过三倍标准差的测定值，称为高度异常的异常值。

15、方差

是衡量随机变量或一组数据时离散程度的度量。概率论中方差用来度量随机变量和其数学期望（即均值）之间的偏离程度。统计中的方差（样本方差）是每个样本值与全体样本值的平均数之差的平方值的平均数。在许多实际问题中，研究方差即偏离程度有着重要意义。方差是衡量源数据和期望值相差的度量值。

16、标准差

又常称均方差，是离均差平方的算术平均数的平方根，用σ表示。标准差是方差的算术平方根。标准差能反映一个数据集的离散程度。平均数相同的两组数据，标准差未必相同。

17、皮尔森相关系数

皮尔森相关系数是用来反映两个变量线性相关程度的统计量。相关系数用r表示，其中n为样本量，分别为两个变量的观测值和均值。r描述的是两个变量间线性相关强弱的程度。r的绝对值越大表明相关性越强。

18、PV（Page View）页面浏览量

指某段时间内访问网站或某一页面的用户的总数量，通常用来衡量一篇文章或一次活动带来的流量效果，也是评价网站日常流量数据的重要指标。PV可重复累计，以用户访问网站作为统计依据，用户每刷新一次即重新计算一次。

19、UV（Unique Visitor）独立访客

指来到网站或页面的用户总数，这个用户是独立的，同一用户不同时段访问网站只算作一个独立访客，不会重复累计，通常以PC端的Cookie数量作为统计依据。

20、Visit 访问

指用户通过外部链接来到网站，从用户来到网站到用户在浏览器中关闭页面，这一过程算作一次访问。

21、Bounce Rate 跳出率

指用户通过链接来到网站，在当前页面没有任何交互就离开网站的行为，这就算作此页面增加了一个“跳出”，跳出率一般针对网站的某个页面而言。

跳出率=在这个页面跳出的用户数/PV

22、退出率

一般针对某个页面而言。指用户访问某网站的某个页面之后，从浏览器中将与此网站相关的所有页面全部关闭，就算此页面增加了一个“退出“。

退出率=在这个页面退出的用户数/PV

23、Click 点击

一般针对付费广告而言，指用户点击某个链接、页面、

banner的次数，可重复累计。比如我在PC端看到一则新闻链接点进去看了一会就关了，过了一会又点进去看了一遍，这就算我为这篇新闻贡献两次点击。

24、avr.time 平均停留时长

指某个页面被用户访问，在页面停留时长的平均值，通常用来衡量一个页面内容的质量。

avr.time=访客数量/用户总停留时长

25、CTR 点击率

指某个广告、Banner、URL被点击的次数和被浏览的总次数的比值。一般用来考核广告投放的引流效果。

CTR=点击数（click）/被用户看到的次数

26、Conversion rate 转化率

指用户完成设定的转化环节的次数和总会话人数的百分比，通常用来评价一个转化环节的好坏，如果转化率较低则需优化该转化环节。转化率=转化会话数/总会话数

27、漏斗

通常指产生目标转化前的明确流程，比如在淘宝购物，从点击商品链接到查看详情页，再到查看顾客评价、领取商家优惠券，再到填写地址、付款，每个环节都有可能流失用户，这就要求商家必须做好每一个转化环节，漏斗是评价转化环节优劣的指标。

28、投资回报率（ROI：Return On Investment ）

反映投入和产出的关系，衡量我这个投资值不值得，能给到我多少价值的东西（非单单的利润），这个是站在投资的角度或长远生意上看的。

其计算公式为：投资回报率（ROI）=年利润或年均利润/投资总额×100%，通常用于评估企业对于某项活动的价值，ROI高表示该项目价值高。

29、重复购买率

指消费者在网站中的重复购买次数

30、流失分析（Churn Analysis/Attrition Analysis）

描述哪些顾客可能停止使用公司的产品/业务，以及识别哪些顾客的流失会带来最大损失。流失分析的结果用于为可能要流失的顾客准备新的优惠。

31、顾客细分&画像（Customer Segmentation & Profiling）

根据现有的顾客数据，将特征、行为相似的顾客归类分组。描述和比较各组。

32、顾客的生命周期价值 (Lifetime Value, LTV)

顾客在他/她的一生中为一个公司产生的预期折算利润。

33、购物篮分析（Market Basket Analysis）

识别在交易中经常同时出现的商品组合或服务组合，例如经常被一起购买的产品。此类分析的结果被用于推荐附加商品，为陈列商品的决策提供依据等。

34、实时决策（Real Time Decisioning, RTD）

帮助企业做出实时（近乎无延迟）的最优销售/营销决策。比如，实时决策系统（打分系统）可以通过多种商业规则或模型，在顾客与公司互动的瞬间，对顾客进行评分和排名。

35、留存/顾客留存（Retention / Customer Retention）

指建立后能够长期维持的客户关系的百分比。

36、相关性分析(Correlation analysis)

是一种数据分析方法，用于分析变量之间是否存在正相关，或者负相关。

37、生存分析（Survival Analysis）

估测一名顾客继续使用某业务的时间，或在后续时段流失的可能性。此类信息能让企业判断所要预测时段的顾客留存，并引入合适的忠诚度政策。

38、算法(Algorithms)

可以完成某种数据分析的数学公式。

39、商业智能（Business Intelligence）

分析数据、展示信息以帮助企业的执行者、管理层、其他人员进行更有根据的商业决策的应用、设施、工具、过程。

40、分类分析(Classification analysis)

从数据中获得重要的相关性信息的系统化过程;这类数据也被称为元数据(meta data),是描述数据的数据。

41、聚类分析(Clustering analysis)

它是将相似的对象聚合在一起，每类相似的对象组合成一个聚类(也叫作簇)的过程。这种分析方法的目的在于分析数据间的差异和相似性。

42、对比分析(Comparative analysis)

在非常大的数据集中进行模式匹配时，进行一步步的对比和计算过程得到分析结果。

43、数据分析(Data Analysis)

是指根据分析目的，用适当的分析方法及工具，对数据进行处理与分析，提取有价值的信息，形成有效结论的过程。

44、数据处理（Data Processing）

数据处理是指根据数据分析的目的，将收集到的数据进行加工、整理，形成适合数据分析的样式，它是数据分析前必不可少的阶段。

45、数据挖掘（Data mining）

数据挖掘是通过使用复杂的模式识别技术，从而找到有意义的模式，并得出大量数据的见解。

46、数据清洗(Data cleansing)

对数据进行重新审查和校验的过程，目的在于删除重复信息、纠正存在的错误，并提供数据一致性。

47、数据质量（Data Quality）

有关确保数据可靠性和实用价值的过程和技术。高质量的数据应该忠实体现其背后的事务进程，并能满足在运营、决策、规划中的预期用途。

48、数据建模(Data modelling)

使用数据建模技术来分析数据对象，以此洞悉数据的内在涵义。

49、数据集(Data set)

大量数据的集合。

50、判别分析(Discriminant analysis)

将数据分类，按不同的分类方式，可将数据分配到不同的群组，类别或者目录。是一种统计分析法，可以对数据中某些群组或集群的已知信息进行分析，并从中获取分类规则。

51、探索性分析(Exploratory analysis)

在没有标准的流程或方法的情况下从数据中发掘模式。是一种发掘数据和数据集主要特性的一种方法。

52、机器学习(Machine learning)

人工智能的一部分，指的是机器能够从它们所完成的任务中进行自我学习，通过长期的累积实现自我改进。

53、网络分析(Network analysis)

分析网络或图论中节点间的关系，即分析网络中节点间的连接和强度关系。

54、异常值检测(Outlier detection)

异常值是指严重偏离一个数据集或一个数据组合总平均值的对象，该对象与数据集中的其他它相去甚远，因此，异常值的出现意味着系统发生问题，需要对此另加分析。

55、模式识别(Pattern Recognition)

通过算法来识别数据中的模式，并对同一数据源中的新数据作出预测

56、预测分析(Predictive analysis)

大数据分析方法中最有价值的一种分析方法，这种方法有助于预测个人未来(近期)的行为，例如某人很可能会买某些商品，可能会访问某些网站，做某些事情或者产生某种行为。通过使用各种不同的数据集，例如历史数据，事务数据，社交数据，或者客户的个人信息数据，来识别风险和机遇。

57、回归分析(Regression analysis)

确定两个变量间的依赖关系。这种方法假设两个变量之间存在单向的因果关系(译者注：自变量，因变量，二者不可互换)。

58、路径分析(Routing analysis)

针对某种运输方法通过使用多种不同的变量分析从而找到一条最优路径，以达到降低燃料费用，提高效率的目的。

59、情感分析(Sentiment Analysis)

通过算法分析出人们是如何看待某些话题。

60、SQL

在关系型数据库中，用于检索数据的一种编程语言。

61、时序分析(Time series analysis)

分析在重复测量时间里获得的定义良好的数据。分析的数据必须是良好定义的，并且要取自相同时间间隔的连续时间点。

62、文本挖掘（Text Mining）

对包含自然语言的数据的分析。对源数据中词语和短语进行统计计算，以便用数学术语表达文本结构，之后用传统数据挖掘技术分析文本结构。

63、可视化(Visualization)

只有正确的可视化，原始数据才可被投入使用。这里的“可视化”并非普通的图型或饼图，可视化指是的复杂的图表，图表中包含大量的数据信息，但可以被很容易地理解和阅读。

64、仪表板(Dashboard)

使用算法分析数据，并将结用图表方式显示于仪表板中。

# / 中国数据分析行业会员单位 ——数据分析师事务所 /

编辑 / 协会会员处 李苗苗 日期 / 2021-02

**数据分析师事务所介绍**

数据分析师事务所（以下称“事务所”）是我国继作为第三方独立服务行业的会计师事务所、律师事务所之后，又一全新的服务行业。事务所是经中数委审批同意、工商局注册备案并具备中国数据分析行业从业资格的单位，受中数委的监督和管理。

事务所由持有CPDA证书的专业数据分析师组成，可为企事业单位提供与大数据相关的专业服务，主要业务范围包括：数据分析与咨询服务、定制化的数据分析产品与大数据平台开发服务、大数据咨询及解决方案服务等。

**数据分析师事务所发展前景**

目前全国有数十家专业的数据分析师事务所遍布在各省市，业务基本围绕着数据的深度分析、业务场景构建、深层次的咨询，以帮助企业实现数据变现的能力。这几年随着广大政府和企业对大数据的重视，数据会越来越多，技术门槛会越来越低，但是大量的数据进行深层次的分析就成为企业竞争的核心、成为企业大数据变现的核心，这对全国的数据分析师事务所来说是一个商业价值巨大的朝阳行业。

**创办数据分析师事务所可享受的协会福利**

1、 提供品牌宣传机会；

2、 享受一年不少于4次实战案例学习，帮助事务所不断提升大数据专业能力；

3、 享受一年一度执业教育免费培训；

4、 帮助事务所解决数据分析师人才不足问题；

5、 给予事务所资质在中国数据分析行业官网上备案、查询；

6、 享受不定期举办免费且形式各异的创业指导活动，有效指导事务所的长期发展；

7、 优先参加我会线上、线下具有一定影响力的活动；

8、 提供协会自主研发的Datahoop智能大数据分析平台最高使用权限。

**数据分析师事务所主要业务范围**

1、 数据分析与咨询服务；

2、 定制化的数据分析产品服务；

3、 大数据平台开发服务；

4、 大数据咨询及解决方案服务；

5、其他。

**申请数据分析师事务所条件**

1、 法人是CPDA数据分析师,成立事务所需要CPDA持证人数不少于3人，法人不是CPDA数据分析师，成立事务所需要CPDA持证人数不少于5人；

2、 注册资金不低于100万;

3、 有一定企业经营经验；

4、 有成熟的业务资源；

5、办公地点不低于100平米。

CPDA® 数据分析师
Certified Projects Data Analyst

# 不甘平凡

UNWILLING
ORDINARY

关注CPDA数据说

ON MY WAY