— 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
2020年第02期 总第42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

危在当下，机在以后；以史为鉴，砥砺前行
——后疫情时代，全力以“复”！

一场突如其来的新冠肺炎疫情搅乱了庚子新春，中国人民经历了一场无比艰苦卓越的抗击新冠病毒的阻击战，每个人都在为这场没有硝烟的战争贡献着自己的力量。终于，在春暖花开之时，我们迎来了国内疫情数据的峰回路转。几个月来，大数据助力全民战“疫”，让我们看到了大数据应用的广阔前景， 特别是在此次新冠肺炎疫情防控工作中，众多城市利用大数据优势，对疫情预警、追踪、监测、宣传，并服务于疫情防控决策和公众参与，对进一步提升政府社会治理能力、公共服务水平和应对大型公共安全突发事件能力均有重大意义。

恩格斯说：“一个聪明的民族，从灾难和错误中学到的东西会比平时多得多。”疫情使得全球众多行业及相关产业链均面临着严峻挑战，而另一方面，疫情为整个社会数字化进程加速带来了新的契机，在线教育、居家办公、线上消费等诸多需求的爆发，开启了数字化转型的新机遇期。疫情过后，如何化危为机，如何在困顿之中寻找新的经济增长点，如何面对逆境快速恢复产业活力并明确适应当前时代发展的新方向，如何利用数字化转型调整商业模式、优化企业组织结构，进而提升企业核心竞争力，成为各国政府、行业、企业需要破解的难题。

疫情危机孕育新机遇，数字中国建设不断向前推动。当前，新一轮科技革命和产业变革加速演进，5G、人工智能、大数据、物联网等新技术、新应用、新业态方兴未艾，为经济的持续增长注入新动能，国家明确提出以“新基建”为核心的全新经济发展方向，已成为提振中国乃至全球经济的一剂强心剂。面对机遇和挑战，数据分析行业也迎来了更加强劲的发展动能和更加广阔的发展空间：在数字经济、技术创新、网络惠民等方面不断取得重大突破，有力推动数字化转型迈上新台阶。随着科技的不断发展，商业智能与数据分析的结合愈加丰富，实现科学化、精准化、高效化，挖掘数据的价值，做出切实可行的决策，对企业的意义非同一般。数字化转型之路也正是众多行业企业转型的必经之路。

后疫情时代，危机不会改变趋势，只是让未来已来。当前中国企业站在数字化转型的重要风口，国家在精准防控疫情的同时积极有序推进复工复产，稳住和支持市场主体，增强回升动力。当前形势下，企业复工复产面临诸多问题，运用大数据与智慧技术助力政府科学决策、优化资源配置、企业有序复工复产的优势被更加凸显出来，对企业复

工复产面临的“痛点”、“难点”和“堵点”，数据分析进一步发挥其自身优势，为政府精准决策、企业经济发展回升持续添砖加瓦。

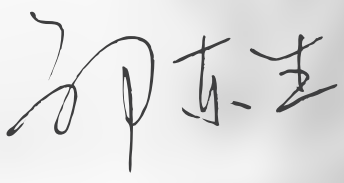
而前不久，在五四青年节到来之际，习近平总书记寄语新时代青年：青春由磨砺而出彩，人生因奋斗而升华。并为广大青年提出目标要求、指明前进方向，激励他们作出更大的贡献。一代人有一代人的长征，一代人有一代人的担当。“长江后浪推前浪”的历史规律，也是“一代更比一代强”的青春责任。后疫情时代，坚定数字化转型布局的方向，抓住数字化经济蓬勃发展的机遇，同时，更应关注和尊重青年思想、认知、成长和价值主张，激励他们适应变化、化危为机、蓄势提能，永做改革创新的奋进者。在此，行业协会也将继续致力于大力培养CPDA数据分析师青年人才，并鼓励、加大力度扶持新青年创立数据分析师事务所，践行以“大数据思维”助力政务数字化发展，企业数字化转型，教育数字化应用，并在数据分析应用价值不断凸显的今天，利用大数据思维为社会数字化转型贡献自己的力量！

志之所趋，无远弗届，穷山距海，不能限也。这是最好的时代，也是最具挑战的时代。大数据的强大赋能，数据分析的决策引导，正前所未有地促进各行各业价值创造方式，如何用创新改变世界，这是今天我们的共同使命。

明者因时而变，知者随事而制，求索之路，始终在传承、坚守中与时俱进，在数据为王的智能时代，让我们共同深耕细作，共同迎接后疫情时期的新机遇！

危在当下，机在以后；以史为鉴，砥砺前行！

中国商业联合会数据分析专业委员会



CPDA®
数据分析师
Certified Projects Data Analyst

后疫情时代，我们起杠！

懂 行业

知 业务

重 分析

善 融通

关注CPDA数据说

我们只培养解决企业关键需求的大数据人才！

咨询热线：400-050-6600

CPDA®
数据分析师
Certified Projects Data Analyst

本期目录 CONTENTS

主办单位
中国商业联合会数据分析专业委员会
编委成员
李苗苗 / 杜天天
出版时间
2020年06月出版 总第42期
美工设计
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
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

/ 数据分析专业委员会被中国商业联合会评为2019年度优秀分支机构 /

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 / 2020-04


荣誉证书

HONORARY CREDENTIAL

数据分析专业委员会：

你分支机构在中国商业联合会2019年度分支机构考核中，被评定为优秀。特颁此证，以示表彰。

中国商业联合会
二〇二〇年三月

/ 江西南昌 | 金融行业线上实战沙龙活动完美收官 /

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 / 2020-04

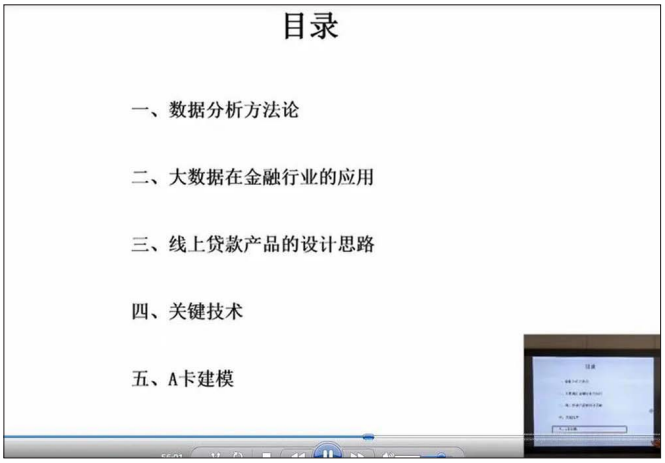
当前，大数据、云计算、移动互联网、人工智能等新一代信息技术飞速发展，正加速与社会各领域深入渗透融合，不断催生着新产品、新模式、新业态；同时，国务院各部门、地方政府以及参与信息化建设的相关单位对大数据应用与发展的需求日益强烈。

为了进一步贯彻落实国家关于大数据应用与发展的相关举措，加快江西省大数据建设发展进程，满足江西地区广大数据分析爱好者、“铁杆粉”们，学习大数据、运用大数据的热情，云上江西携手中国商业联合会数据分析专业委员会于4月6日推出以《线上贷款产品设计思路及关键技术》为主题的线上沙龙活动。

此次活动特邀中国商业联合会数据分析专业委员会特聘专家，《中国大数据人才培养体系标准》专家组成员李军博士为大家带来精彩内容。活动吸引了来自江西地区的银行、保险、证券、租赁、互联网等金融行业近500名数据分析爱好者参加。

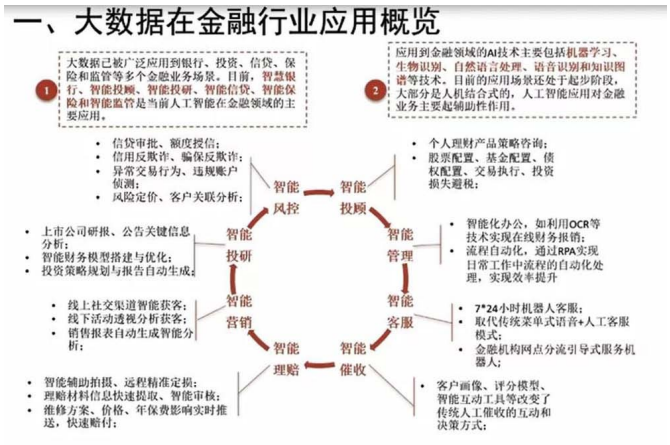
此次活动深度聚焦金融行业的企业级实战案例，从实际应用场景入手，为广大数据分析爱好者量身定制了大数据及数据分析方法，引导其结合自身所在行业与环境分享多样化数据分析解题思路，帮助广大数据分析爱好者从容应对未来挑战。CPDA智慧教室为本次沙龙活动助力，让广大数据分析爱好者身临其境的参与到了此次活动中，与老师在线上进行互动交流、深入探讨学习，体验了一场颠覆式的“云直播”，现在就让我们一起来回顾吧！

线上沙龙回顾一：数据分析方法论



李军老师先从大数据产业链、大数据全生命周期管理、数据分析方法论（PDCA戴明环）、业务和数据、技术三位一体等方面阐述了数据分析方法论。

线上沙龙回顾二：大数据在金融行业的应用



通过招商银行、宁波银行、广州农商行在智能风控、智能客服、智能营销、智能投顾、智能渠道、智慧生活等方面的实际案例，李军老师为大家介绍了大数据在金融行业的一些成熟应用。

线上沙龙回顾三：线上贷款产品的设计思路

李军老师通过分析线上贷款产品——针对市场前景广阔的小微信贷产品的四大痛点（信息不对称程度高、数据缺乏（财务、征信等）、关联性风险强、风控要求与成本效率难平衡）提出数据已成核心驱动力，新技术如客户画像、统计建模、NLP、KG、CV、深度学习为小微信贷产品提供了整体解决方案。

线上沙龙回顾四：关键技术：反欺诈、A/B/C卡

关键技术主要涉及机器学习常用算法、完整的建模过程、数据获取、特征工程、模型构建、反欺诈产业链六个方面，并详细阐述了A卡的建模过程和方法。

线上沙龙回顾五：逻辑回归模型在A卡开发中的应用

李军老师从明确A卡建模目的到详述数据获取的方法步骤以及对逻辑回归（Logistic）建模进行效果评估等内容，让我们了解到逻辑回归模型在A卡开发中的应用内容。对于输出的每一条客户模型预测结果，通过标准化进行模型分数排序，即新打分卡，可以判断出客户的违约风险程度。

近2个小时的精彩内容，兼具理论与实践、讲授与互动，并穿插以李军老师从业心得和经验分享，让所有参加此次线上沙龙活动的小伙伴收获颇丰，对于此次寓意深刻而又富有实战作用的课程，小伙伴们纷纷表示大有收获，并纷纷期待下期的线上沙龙活动。

特邀讲师介绍

管理科学与工程博士、计算机硕士、中科院大学企业导师、高级经济师、信息系统项目管理师。

曾任职国家发改委、中南国农业银行总行副处长，现任某金融集团大数据中心总经理，兼任某银行董事。中国商业联合会数据分析专业委员会特聘专家，《中国大数据人才培养体系标准》专家组成员。 具有20年金融信息化、大数据分析经验，对中小企业的商业模式有一定研究，并主导设计了针对中小企业的营销模式、信贷产品。精通数据在企业应用的方法论，能找准企业在数据应用中的痛点并采取合适的解决方案，精通数据在金融行业的应用。 精通统计机器学习的算法原理及应用，熟悉深度学习的各类算法及应用。

中国商业联合会数据分析专业委员会

中国商业联合会数据分析专业委员会（China Data Analysis Committee ,China General Chamber Of Commerce，缩写CDAC），是以数据分析师事务所等企业为主体，以及从事与数据分析业相关的数据分析、投资服务、数据分析研究等方向的科研院所、大专院校、经营性企业、服务性企业和相关团体与个人自愿组成的全国性行业组织，是经国务院国有资产监督管理委员会审核同意、中华人民共和国民政部正式批准和登记的中国数据分析行业主管协会。

云上（江西）大数据发展有限公司

云上（江西）大数据发展有限公司（以下简称“云上江西”）是经江西省委省政府同意，由江西省创业投资管理有限公司和江西省铁路投资集团有限责任公司共同出资设立，以服务数字政府建设、盘活政府数据资产为目标，引导数字产业发展，推动江西数字经济快速发展。云上江西公司的成立，对于推动数字产业化和产业数字化，培育发展新动能，支撑江西高质量跨越式发展具有重大意义。

/ 4月24日CPDA线上沙龙活动完美收官，全程精彩不断！ /

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 / 2020-04

4月24日，CPDA为全国数据分析爱好者带来了一场精彩的主题为《get数据分析思维模型只要6步》的线上沙龙活动。

本次活动特邀西安交通大学管理学博士、中国商业联合会数据分析专业委员会研发师资李妹老师来进行分享，活动吸引了来自全国各行业、各领域的近600名数据分析爱好者。

在将近2个多小时的活动中，广大数据分析爱好者积极参与到了各环节的互动中，与老师在线上进行互动交流、深入探讨学习。让想了解数据分析的同学们，体验了一把“宅家沙龙”的畅快，现在就让我们一起来回顾吧！

线上沙龙回顾一：数据分析的前沿应用案例

活动开始，李妹老师通过新冠肺炎的疫情模型分析、facebook用户画像精准推送及婚姻离婚率预测等当下热点问题分享作为案例，将同学们带到了数据分析的场景之中，并让大家深刻理解数据分析在各行各业中的实际应用价值，大到政策制定、小到个人决策，一切问题都可以用数据说话。

线上沙龙回顾二：学习数据分析的一些误区

紧接着，李妹老师分享了自己在数据分析工作中的实战

经验结和小秘籍，为大家讲解了在学习数据分析过程中，遇到的一些错误和误区，让大家深刻理解了思维、方法与工具之间的优先级排序，正确、快速地建立数据分析知识体系，少走弯路。

线上沙龙回顾三：数据分析师知识体系构

接下来，为让大家正确地看待数据分析师这个职位。李妹老师以“数据分析师都将被人工智能取代？”的疑问抛出了一个值得大家深思的问题，当然也给出了答案——数据分析的核心竞争力应是思维方式，而不是重复性的抓取数据的工作。人工智能只能解决数据分析工作中最初级的问题，让数据分析师留出更多精力聚焦在分析本身。同时鼓励让大家成为能进行趋势预测、行业研究、评估的数据分析专业人才，最终为企业保驾护航。

线上沙龙回顾四：数据分析思维六步法及其案例解析

最后，李妹老师结合福建移动和香港赌马公司的真实案例，为我们讲解并验证了数据分析思维6步法的实现过程。让大家清晰完整地了解数据分析的思路，真正地通过数据分析解决工作、生活中遇到的各种问题。

近两个小时的精彩分享，兼具理论与实践、讲授与互动，并穿插以李老师从业心得和经验分享，让所有参加此次线上沙龙活动的小伙伴收获颇丰，对于此次寓意深刻而又富有实战作用的沙龙活动，小伙伴们纷纷表示大有收获，并纷纷期待下期的线上沙龙活动。

/ 5月15日CPDA数据分析师线上沙龙活动完美收官，干货满满、千人共享！ /

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 / 2020-05

5月15日，我们为全国数据分析爱好者带来了一场精彩的主题为《疫情之后的经济趋势与企业大数据应用》的线上沙龙活动。

本次活动特邀有着10余年资深数据分析职场导师经验的王兴海老师来进行分享，活动吸引了来自全国各行业、各领域的近千名数据分析爱好者。

在将近1个半小时的活动中，广大数据分析爱好者积极参与到了各环节的互动中，与老师在线上进行互动交流、深入探讨学习。让想了解数据分析的同学们，体验了一把“引爆思维”的满足，现在就让我们一起来回顾吧！

线上沙龙回顾一：当前经济形势分析

沙龙活动开始，王兴海老师采用娓娓道来的方式，带领大家顺着时间逻辑回顾了从1840年到2020年历年的自然灾害，让我们直观地了解到当前严峻的外部环境干扰；通过今年统计局一季度的财政数据，演化出当前萧条的经济形势，并运用宏观经济学GDP模型，阐述了各国及国内在宏观层面的应对措施。

线上沙龙回顾二：企业面对的现实与问题

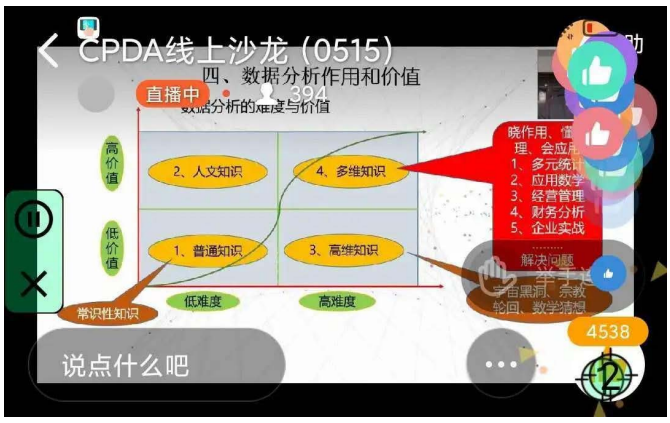
面对严峻的经济前景，中国经济的巨轮仍然具有劈波斩浪、砥砺前行的强大内力。紧接着，王兴海老师从中国经济的长远趋势，到世界经济和中国经济的宏观周期，全面分析了当前及今后一段时间里企业面对的现实与问题，特别是在缺资金、缺市场、缺产品、缺人才四个方面作出了独到的剖析，提出了一些富有前瞻性的理论观点。

线上沙龙回顾三：我们的机会和应对策略

有困难，就会有机会。疫情面前，我们该如何抓住数据化转型的契机为疫情后的企业续命？王兴海老师从数字经济模式下的产业投资、5G到NG的科技投入、DT时代的一切皆可数字化三个层面强调了数据化转型对于现代企业的重要性，在资金、人才、产品有限的情况下如何利用数据化策略使企业立于不败之地。

线上沙龙回顾四：数据分析作用和价值

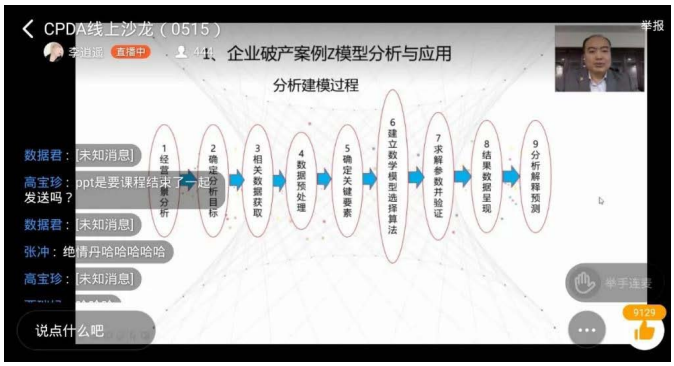
接下来，王兴海老师通过B-17轰炸机与朝鲜战争两个生动的故事证明了用数据说话的重要性。更重要的是，他告诉了我们数据决策力已经变成一种不可或缺的能力。



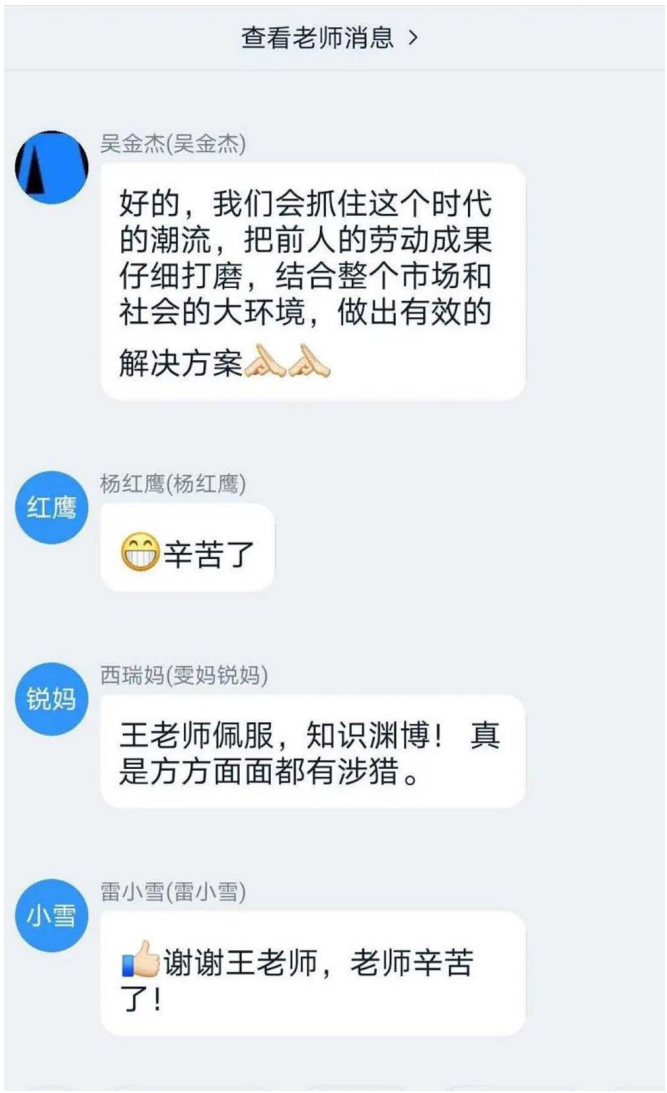
预测企业未来、量化决策。杜绝“拍脑门”决策，利用大数据支撑决策，精准获客，降本增效。对企业的销售、资金、成本、利润等情况进行预测，合理应对未来的不确定，给决策者提供量化决策依据。

线上沙龙回顾五：大数据视野下，企业经营中如何进行应用创新

王兴海老师接着从事物的相关、因果到敏感性分析，展开细致的数学建模思路和流程，利用定性分析和定量分析的方法解决数据分析的常见场景。并以某制造业公司破产的案例，通过数据分析，建立数据模型，让我们看清企业的风险所在，以及规避破产的防范措施及流程。



最后，王兴海老师就大家关心的个性化问题进行了现场答疑，小伙伴们纷纷表示大有收获，本次沙龙活动在大家的一致认可下圆满结束！



/ 农业农村部办公厅关于印发《2020年农业农村部网络安全和信息化工作要点》的通知 /

来源 / 农业农村部办公厅　编辑 / 协会会员处 李苗苗　日期 / 2020-05

【导读】5月8日，农业农村部办公厅印发了《2020年农业农村部网络安全和信息化工作要点》，要点指出要紧紧围绕实施乡村振兴战略、数字乡村战略总体要求，大力实施数字农业农村建设，深入推进农业数字化转型，扎实推动农业农村大数据建设，全面提升农业农村信息化水平，用信息化引领驱动农业农村现代化，为全面建成小康社会提供有力支撑。

农业农村部办公厅关于印发《2020年农业农村部网络安全和信息化工作要点》的通知

各省、自治区、直辖市农业农村(农牧)厅(局、委)，新疆生产建设兵团农业农村局，部机关各司局、派出机构、各直属单位：

为贯彻落实《中共中央、国务院关于抓好“三农”领域重点工作确保如期实现全面小康的意见》《数字乡村发展战略纲要》要求，做好2020年农业农村信息化重点工作，我部研究制定了《2020年农业农村部网络安全和信息化工作要点》。现印发你们，请结合实际，认真组织实施。

农业农村部办公厅

2020年5月7日

2020年农业农村部网络安全和信息化工作要点

2020年农业农村部网络安全和信息化工作的总体要求是，以习近平新时代中国特色社会主义思想为指导，深入贯彻党的十九大和十九届二中、三中、四中全会以及中央农村工作会议精神，认真落实全国农业农村厅局长会议决策部署，紧紧围绕实施乡村振兴战略、数字乡村战略总体要求，大力实施数字农业农村建设，深入推进农业数字化转型，扎实推动农业农村大数据建设，全面提升农业农村信息化水平，用信息化引领驱动农业农村现代化，为全面建成小康社会提供有力支撑。

一、大力实施数字农业农村建设

1.加快实施数字乡村建设。贯彻落实《数字乡村发展战略纲要》部署，围绕数字乡村建设重点任务，强化工作举措，扎实推进各项工作。指导各级农业农村部门推动实施《数字农业

农村发展规划（2019—2025年）》。（部市场司、规划司）

2.深入推进信息进村入户工程。强化工作指导与建设考核，持续推进益农信息社建设。依托益农信息社，结合“互联网+”农产品出村进城工程，推动农产品上行，助力农民脱贫致富的同时提高益农信息社持续运行能力。推动各类服务资源通过益农信息社向农村下沉，充分利用智能终端等新型手段，创新推动线上线下结合的服务方式。发挥益农信息社信息采集“传感器”作用，探索信息直报机制模式。（部市场司）

3.加快启动实施“互联网+”农产品出村进城工程。研究出台试点工作规范，指导地方开展工作，加快推动工程落地实施。优先选择包括贫困地区在内的100个县开展试点，建立试点工作成效评估机制，加强督促检查和跟踪评估。探索建立数据采集平台，推进各试点县、各参与主体与平台对接。（部市场司）

4.抓好数字农业农村试点工作。推进国家数字农业农村创新中心和重要农产品全产业链大数据建设，进一步优化数字农业试点县项目布局和设计，加强项目建设过程管理和指导，加快构建农业农村基础数据资源体系，推动农业农村数字化转型，探索可复制、可推广的数字农业农村建设模式。（部市场司、规划司）会同中央网信办开展数字乡村试点县建设，加强统筹规划，推动农业农村数字化资源集聚共享，积极探索典型经验做法，及时总结推广有益经验。（部市场司）继续举办数字农业发展培训班。（部规划司）

5.持续推动乡村治理信息化。鼓励指导各地开展多种形式的数字乡村治理实践，引导社会力量推进乡村治理信息化建设。指导各地加强农村财务会计、农民负担监管、农业社会化服务以及农村宅基地管理等工作信息化水平。（部合作经济司）建立健全农村社会事业监测体系，在农村人居环境等领域开展调度与监测。（部社会事业司）

6.推进农村集体资产管理信息化。加快推动农村集体资产大数据资源共享，建设农村集体资产监管平台，提升农村集体资产监管水平。指导各地完善土地承包管理信息数据库，促进数据互联互通。推动各地探索开展承包合同、流转合同网签等信息化管理手段。（部政策改革司）

二、深入推进农业数字化转型

7.推进农产品质量安全信息化。积极推进大数据、云计算等现代信息技术与农产品质量安全工作全面融合，探索构建“智慧农安”，开展智慧监管试点示范。（部监管司）

8.推进种植业信息化。加快推进种植业信息资源整合，持续完善农情、灾情、病虫情和土壤墒情以及农药、肥料等生产资料信息化监测管理手段。（部种植业司）

9.推进畜牧兽医信息化。加快推进部级畜牧兽医政务信息资源交换共享，促进数据互联互通。全面推广养殖场直联直报。开展强制免疫疫苗财政资金直补和生猪全产业链监管监测信息化试点工作。继续推动奶业主产省数字奶业建设。（部畜牧兽医局）

10.推进渔业渔政信息化。加快各地海洋渔船通导安全装备项目实施，新建一批渔业无线电岸台基站，配备一批渔船安全通导终端，完成全国渔船渔港动态管理系统容灾备份中心建设，做好运行保障工作。推动水产品市场流通、渔船渔港监管、渔政执法、渔业资源环境监测等领域信息化建设，加强网络安全管理。（部渔业渔政局）

11.推进种业信息化。完善数字种业建设规划，推动种业创新发展、提升种业公共服务效能。加快种业政务和业务在线化建设，打通各业务板块数据信息，加快推动种子品种、质量和市场主体可追溯。提升种业高阶智能分析水平。（部种业司）

12.推进机械化与信息化融合。大力推行“互联网+”农机管理服务，持续推进农机鉴定、认证、购置补贴、安全监理、统计等数据资源互联互通。（部农机化司）

13.推进农田建设信息化。推动全国农田建设综合监测监管系统建设，将农田建设项目纳入监测监管范围，及时掌握各地建设进度。加快推进高标准农田上图入库工作，构建全国农田建设“一张图”。利用现代空间信息技术，试点开展高标准农田建设、利用、管护等监测。（部农田建设司）

14.加强农业农村服务信息化普及。丰富完善“中国农技推广”“云上智农”等功能和内容，进一步整合汇聚资源，全面提升基层农技推广和高素质农民培训的服务能力与信息化水平。（部科教司）持续开展农民手机应用技能培训，采用线上线下相结合的形式，培训提升农民手机应用能力，不断完善线上培训功能、丰富培训资源、扩大师资队伍和培训受众面。（部市场司）继续做好重点面向“三区三州”等深度贫困地区的农村实用人才带头人农业农村电子商务专题培训班。（部信息中心）

三、扎实推动农业农村大数据建设

15.依托现有资源建设农业农村大数据中心。加快建设农业农村大数据基础设施，完善数据存储运算条件环境。加速推进农业农村数据信息的汇聚共享和分析利用，建设农业农村大数据平台，实现数据可用可查可视和科学高效管理，加大数据精准采集、行业监测、动态预警、建模分析、决策辅助、共用共享、综合展示力度，深化强化大数据技术在农业农村生产、经营、管理等领域的应用，为农业农村经济和社会高质量发展提供坚实的数据支撑。（部市场司、办公厅、规划司、计财司、科教司、信息中心、中国农业科学院）

16.加快推进重要农产品全产业链大数据建设。聚焦重要农产品品种，打通产业链各环节核心数据的采集、分析和应用，加快关键共性技术研发、挖掘利用、关联分析和成果应用，探索完善政府部门与社会各方数据资源共享共建共用机制。加大重要农产品全产业链大数据中心建设，促进数据共建共享，实现与全国农业农村大数据平台的统一对接。（部市场司、规划司、信息中心）

17.加大数据开放共享服务力度。改版上线农业农村部网

站数据频道，丰富数据资源，促进数据开发利用，提供一站式数据服务。不断提升农兽药基础数据平台、国家农产品质量安全追溯管理信息平台、重点农产品市场信息平台和新型经营主体信息直报系统的服务能力，优化完善“四平台”功能作用，加强数据共享共用。完善农业利用外资和对外投资信息采集体系，加强中国农产品供需分析系统推广应用。（部种植业司、畜牧兽医局、监管司、市场司、计财司、国际司、信息中心分头负责）

四、进一步夯实农业农村信息化工作基础

18.做好农业农村信息化延伸绩效考核。依据农业农村部绩效管理办法、专项工作延伸绩效管理实施办法，结合各省（市）现有工作基础，开展农业农村信息化发展延伸绩效管理试点工作。（部市场司）

19.推进农业信息化标准化工作。宣贯《农业信息化标准体系（暂行）》，进一步发挥其指导、规范、引领和保障农业信息化建设与发展的作用。加强农业信息化标准化管理，推动重点领域标准建设。（部市场司、信息中心）

20.持续开展农业农村信息化发展水平监测与评价工作。组织开展农业农村信息化发展水平评价工作，进一步优化评价指标体系、省县两级数据采集系统，加强数据整理、分析、开放等功能，完成2020年评估报告。（部市场司、信息中心）

21.加强农业农村信息化发展经验总结推广。加快推进《中国农业百科全书·农业信息化卷》编撰工作。联合中央网信办，编制《数字乡村发展报告（2020）》等报告。举办数字农业农村论坛，组织各类社会主体，参与农业农村信息化展览展示、论坛等交流活动，营造全社会广泛关注、共同参与数字乡村建设的氛围，形成推进合力。（部市场司、信息中心）

/ 两会 | 大数据产业成两会热点，10余位代表、委员建言！ /

来源 / 山东省大数据局　编辑 / 协会会员处 李苗苗　日期 / 2020-05

受疫情影响，2020年全国两会不仅延迟召开，持续时间也比原计划缩短了4天半。

今年的两会时间虽短，但来自全国各地的代表们依然踊跃建言献策，本文搜集了10余位代表有关“大数据”的提案议案，发现“共享”“安全”“立法”成为了代表们建议中的关键词。

全国人大代表杨帆：抓大数据促智能化加快建立全国一体化的数据“聚通用”体制机制

全国人大代表、重庆市大数据应用发展管理局副局长杨帆表示，抓大数据促智能化是增强发展新动能、推动高质量发展的必然选择，因此她建议，加快建立全国一体化的数据“聚通用”体制机制，加强对地方大数据管理部门的统一指导，统筹推进全国各级各部门的数据共享开放、融合应用、流通交易、新型智慧城市建设等工作，进一步打破数据壁垒，特别是解决国家部委“数据”无法回归省市、跨省市数据共享与业务互联困难等问题。

全国人大代表杨剑宇：搭建大数据共享平台满足各领域需求

全国人大代表、中国移动通信集团河南有限公司总经理杨剑宇杨剑宇表示，目前对数据资源的管理，还处在“谁采集、谁占有、谁管理”的阶段，集中化程度不高，公共大数据内容分散，数据共享存在行业壁垒，数据浪费和流失问题也时有出现。杨剑宇建议主管部门搭建大数据共享交换平台，建立数据交换共享机制；政府加强大数据的安全保障，进一步完善法律法规，确保大数据在数据归集、存储、应用等各个环节的安全可靠运行。

全国人大代表陈建华：构建雄安新区金融大数据中心

全国人大代表、人民银行石家庄中心支行行长陈建华表示，当前金融大数据搜集、整理与应用能力已成为主权大国之间开展金融竞争的主要依托力量，金融大数据治理也已成为全球金融治理的重要内容并成为推动全球金融治理机制变革和演化的重要推动力量。因此，从保障雄安新区定位、实现现代金融监管、实现现代金融监管等角度来看，陈建华建议构建雄安新区金融大数据中心。在构建“金融大数据中心”的具体措施方面，陈建华建议，第一要明确金融大数据中心功能定位；第二要构建金融大数据中心发展机制；第三要推进金融大数据中心技术应用；第四要出台大数据相关法律政策，从立法层面出台“大数据法”。

全国人大代表孙丕恕：加快健康医疗大数据产业化发展

全国人大代表、浪潮集团董事长兼CEO孙丕恕表示，在国家各类相关政策、社会广泛需求和技术创新等因素影响下，健康医疗大数据正加快在临床科研、新医药创制、全民健康管理、互联网+医疗服务、疾病预警和公共卫生监测等领域的深度应用。为更好地统筹推进健康医疗大数据建设应用，孙丕恕在今年《关于加快健康医疗大数据产业化发展，助力健康事业和数字经济发展的建议》中提出，要加快开展系统的健康医疗大数据产业发展的法律法规和政策体系建设，在确保数据安全和隐私保护前提下，完善健康医疗大数据授权运营机制，加快健康医疗大数据在药械企业、保险行业的应用和发展，营造“以数引智、以数育商”的产业发展氛围。

全国人大代表魏明：建议加快制定“数据安全法”

全国人大代表、广东移动董事长、总经理魏明今年发起并由数十名代表联署的关于加快制定“数据安全法”的议案指出，伴随互联网的高速发展及物联网、工业互联网、云计算、大数据等新兴领域和新兴技术的快速崛起，数据安全不仅面临数据窃取、篡改与伪造等传统威胁，同时也存在日益增多的数据滥用、个人信息与隐私泄露等新问题。魏明表示，加快制定数据安全法已刻不容缓，并提出了确立数据主权、明确数据安全法的管辖范围，对数据经营进行牌照化管理，建立数据采集、加工和利用业务的准入制度，完善数据安全监管体系和数据安全监测预警、应急处置机制，建立责任主体问责制度等一系列建议。

全国人大代表俞光耀：为数据质量立法

全国政协委员、上海申通地铁集团有限公司董事长俞光耀表示，国内缺乏统一的数据质量标准，因此俞光耀建议以《促进大数据发展行动纲要》《大数据产业发展规划（2016-2020年）》等国家指导性文件、地方性法规为基础，建议国家层面加快研究建立覆盖数据采集、整理、应用和隐私保护等全流程的数据质量监管法制体系。为企业获取和加工数据、社会利用数据，以及政府部门进行相应的监管活动提供法制保障。

全国人大代表张爱军：大数据助力中小微企业发展

全国人大代表、江苏省宿迁市委书记张爱军认为，在毫不松懈地继续抓细抓实疫情防控的同时，当务之急是抓紧迅速推进经济社会步入正常轨道。张爱军代表建议，中央相关部委抓紧运用大数据、信息化等手段助力产业特别是中小微企业发展。

全国人大代表檀结庆：扩大国产数据库金融领域试点

围绕“加快金融市场基础设施建设，稳步推进金融业关键信息基础设施国产化”，全国人大代表、合肥工业大学应用数学研究所所长檀结庆提出三点建议：扩大国产数据库金融领域试点，设立数据库产业引导基金，搭建前沿技术信息共享平台。

全国人大代表石蓉：立法解决大数据共享障碍

全国人大代表、贵阳市公安局刑侦支队副支队长石蓉认为，现在国家、省、市数据共享体系初步形成，但是没有彻底打通到省、市、县级，使得基层缺乏常态化的数据融合共享制度保障，各行业间依然存在数据共享融合壁垒，如一些垂直管理系统由于条线上的数据开放权限限制，数据更新不及时，没有真正做到实时、深度、全面融合。石蓉建议，制定相关法律，完善规范体系，彻底解决部门之间、行业之间大数据共享障碍问题。

全国政协委员王均金：建立全国农民工信息大数据平台

全国政协委员、均瑶集团董事长王均金指出，现有的农民工大数据相关平台存在范围不广、字段不够、积累不足等问题，未真正满足“大数据”的概念，在精准分析匹配方面也未充分发挥作用。对此，他建议，建立全国性的农民工信息大数据平台。利用大数据技术，构建信息实名、数据完备的全国性农民工信息大数据平台。建立农民工真实信息档

案，记录个人的家庭情况，就业经历、培训历史、社保、健康状况、目前就业情况等信息，并实时更新。同时，发挥大数据平台的数据分析能力，为农民工提供精准的岗位引导服务，提供切实有效的就业支持，实现农民工、企业、技能培训机构之间的“三重对接”。

全国政协委员徐晓兰：加大工业互联网大数据中心投资强化数据共享?

全国政协委员、中国工业互联网研究院院长徐晓兰表示，当前，我国工业互联网数据资源总量呈爆炸性增长，但是各地区各行业的数据资源间仍存在孤立、分散、封闭等问题，数据价值未能得到有效利用。我国已成立国家级工业互联网大数据中心，但实现对工业大数据资源的统一管理和使用，需要构建跨层级、跨地区、跨行业的国家工业互联网大数据中心体系，以彻底解决数据“孤岛”问题。对此，她建议，构建完善的工业互联网数据合作共享机制；强化工业互联网数据合作共享生态；加大对工业互联网大数据中心建设的投资力度。

全国政协委员童国华：完善大数据体系建设助推国家治理体系和治理能力现代化

全国政协委员、中国信科集团党委书记、董事长童国华建议，建设国家大数据中心，为高质量发展赋能；持续完善数字政府建设，夯实数字化支撑能力；以构建公共应急体系为切入点，释放大数据价值；探索创新公共服务模式，提升群众获得感；优化社会信任体系，强化政府公信力；解决困扰大数据运用难题，共同营造信息充分可控、安全运行的绿色生态。

全国政协委员程红：构建青少年体质监测预警数据共享体系

全国政协委员、北京市政协副主席程红建议构建青少年体质监测预警数据共享体系。打破部门壁垒，整合婴幼儿保健记录、学校体检数据、医院就诊信息、社区健康记录等，形成完整的个人健康数据链。构建全国儿童青少年体质健康大数据平台，完善体制健康监测等，定期公布各地儿童青少年健康数据排名，实现由重治疗向重预防转变，同时加强对卫生健康教育评测与干预的理论研究、科普工作和人才培养。

/ 银保监会启动专项治理摸底金融机构数据质量 /

来源 / 上海证券报　编辑 / 协会会员处 李苗苗　日期 / 2020-05

银行保险行业将启动专项治理行动，对金融机构数据质量进行摸底。银保监会于近日向各银保监局、会管金融机构下发的《关于开展监管数据质量专项治理工作的通知》中，透露了这一重磅消息。

近年来，金融机构在业务快速发展过程中，积累了客户数据、交易数据、外部数据等海量数据，数据已成为金融机构的重要资产和核心竞争力。

数据治理也一直是监管部门的重点工作。近段时间以来，多家银行、保险公司因数据质量及数据报送存在违法违规行为，而领到监管罚单，并被通报批评，其中不乏大型金融企业。这些处罚也折射出机构在对待数据质量及报送问题上的不严谨，主要表现为：数据准确性和完整性欠缺，时效性和适应性不足。这一方面阻碍了金融机构向高质量方向发展，另一方面也影响了监管效率。

数据治理亟待加强，数据质量亟待提升。据了解，银保监会将成立监管数据质量治理工作领导小组，由分管会领导任组长，由此可见监管对这次治理行动的重视程度。此次治理的数据范围包括监管数据及相关源头数据，其中监管数据是指按照监管要求定期报送银保监会及其派出机构的监管统计数据和其他监管数据。

根据同步下发的专项治理方案，此次治理重点关注四大数据质量，包括数据真实性、准确性、完整性、及时性等。机构范围主要包括：大型银行、股份制银行、城市商业银行、农村商业银行、外资银行、信用合作社等吸收公众存款的金融机构，政策性银行，国家开发银行以及保险集团（控股）公司、保险公司、保险资产管理公司。银保监会及其派出机构依法监管的其他银行保险机构参照执行。

根据通知要求，此次专项治理工作要压实监管数据质量责任，以监管数据质量问题为导向，通过机构自查自评和监管检查评估双向驱动，促进银行保险机构在发现问题、分析原因、落实整改的过程中，不断提升监管数据质量。

银行保险机构应高度重视监管数据质量问题和薄弱治理环节，对短期内能够解决的，立查立改；对暂时无法解决的，确保按计划逐步整改到位。要从发现的问题出发，追根溯源，强化源头治理，打牢数据质量根基；夯实管理基础，补齐组织、制度、机制、系统等方面的工作短板，建立全面提升监管

数据质量的长效机制。

各银保监局要督促银行保险机构切实落实整改措施，同时完善监管机制、流程和手段，持续推动银行保险机构提高监管数据质量。各银保监局要坚持定期开展监管数据质量通报，形成常态化监督机制，对数据差错问题严重、屡错不纠、整改不力、治理责任履行不到位的机构，综合运用监管约谈、通报批评、责令整改、行政处罚、挂钩监管评级等措施，加大问责与约束力度，督促银行保险机构牢固树立底线意识，切实落实监管数据质量责任。

根据工作进度安排，此次专项治理行动将历时1年，主要分为五步：2020年5月为工作启动阶段，2020年6月至8月为银行保险机构自查自评阶段，2020年9月至12月为监管检查评估阶段，2021年1月至4月为问题整改阶段，2021年5月为总结交流阶段。

在总结交流阶段，银保监局要针对辖内监管数据质量专项治理情况，形成工作总结报告。银行保险机构则要总结专项治理工作，对监管相关数据质量及其治理情况进行内部考评。

/ 贵阳市大数据发展工作领导小组办公室 关于印发《2020年全市大数据发展工作要点》的通知 /

来源 / 贵阳市大数据发展工作领导小组办公室　编辑 / 协会会员处 李苗苗　日期 / 2020-04

各区、市、县人民政府，高新技术开发区、经济技术开发区、贵阳综合保税区、贵州双龙航空港经济区管委会，市有关部门，市大数据发展工作领导小组各成员单位：

现将《2020年全市大数据发展工作要点》印发给你们，请认真抓好贯彻落实。

附件：《2020年全市大数据发展工作要点》

贵阳市大数据发展工作领导小组办公室

2020年4月13日

2020年全市大数据发展工作要点

2020年全市大数据工作的总体要求是：坚持以习近平新时代中国特色社会主义思想为指导，深入贯彻落实习近平总书记对贵州工作重要指示精神，致2018、2019数博会贺信精神和关于国家大数据战略重要论述，坚定不移实施大数据战略行动，坚持“四个强化”、做实“四个融合”，持续深耕五大领域，加速构建五大体系，全面建成数博大道核心区，以高标准要求服务我市高水平开放、高质量发展，全力做大数字产业。

一、进一步加快数字产业发展

1.加快推动软件和信息技术服务业发展。以贵阳市信息技术服务产业集群获批国家战略性新兴产业集群为契机，加速发展软件和信息技术服务产业。2020年，实现软件和信息技术服务业软件业务收入突破200亿元，增长25%；规上互联网和相关、软件和信息技术服务业营收突破100亿元，增长22%。新增规上互联网和相关、软件和信息技术服务业企业20家。（牵头单位：市大数据局、市发展改革委、各区〔市、县〕政府、各开发区管委会）

2.加快推动产业集聚发展。充分发挥数博大道承载作用，打造5个10亿级市级大数据（信息技术服务业）产业集聚示范基地，推动产业集聚发展，提升大数据产业发展竞争力。（牵头单位：市大数据局、市发展改革委、云岩区政府、南明区政府、白云区政府、观山湖区政府、高新开发区管委会、经济技术开发区管委会）

3.加快推动软件服务外包提档升级。以申建国家服务外包示范城市为契机，加快发展以软件和信息技术服务、大数据服务为重点的软件服务外包产业，打造“贵阳服务”品牌。2020年，软件外包服务收入突破90亿元，占软件业务收入达45%。（牵头单位：市大数据局、市商务局、各区〔市、县〕政府、各开发区管委会）

4.深入实施“百企引领”行动。立足大数据五新领域创新发展，着力提升大数据企业竞争力。2020年，培育100户软件和信息技术服务业及大数据产业引领企业。（牵头单位：市大数据局、各区〔市、县〕政府、各开发区管委会）

5.深入推进招商及重大项目建设。加强市区联动，加快推动腾讯云产业生态项目、浪潮大数据产业基地、欧比特卫星大数据综合应用示范基地等重大项目建设，着力推进华为鲲鹏生

态基地、中智国际人才服务等重大项目落地。2020年，引进落户数字经济骨干企业8家、大数据优强企业60家、“寻苗行动”潜力企业80家。（牵头单位：市大数据局、市工业和信息化局、市投资促进局、各区〔市、县〕政府、各开发区管委会）

6.加快推动大数据安全产业发展。加快推进贵阳大数据安全示范区建设，2020年完成投资3亿元，新增建设面积5万平米，营收突破1亿元。以“数字孪生城市靶场”为抓手，加快国家大数据安全靶场（二期）建设。构建大数据、网络安全产品测评及安全技术认证体系。办好2020贵阳大数据及网络安全精英对抗演练。（牵头单位：经济技术开发区管委会、市大数据局、市网信办、市公安局）

7.助力企业平稳发展。实施《贵阳市促进软件和信息技术服务业发展的若干措施》，加大财政资金扶持及引导作用，助力企业渡过疫情难关，实现平稳发展。（牵头单位：市大数据局、市财政局、各区〔市、县〕政府、各开发区管委会）

8.做好产业发展顶层设计。谋划贵阳市贵安新区大数据协同发展。启动贵阳市“十四五”大数据发展专项规划、贵阳市“十四五”软件和信息技术服务业发展专项规划编制工作。出台《贵阳市加快发展软件服务外包产业行动计划（2020-2022）》。（牵头单位：市大数据局、市商务局）

二、进一步深化大数据融合应用

9.深入实施“万企融合”行动。以工业互联网标识解析二级节点（电子行业应用服务平台）等项目建设为契机，加快推进贵阳智能制造协同共享平台建设。加快推动工业、农业、服务业数字化改造，建设30个省级融合标杆项目、实施300个市级融合示范项目、带动600户企业与大数据深度融合。（牵头单位：市大数据局、市发展改革委、市工业和信息化局、市农业农村局、各区〔市、县〕政府、各开发区管委会）

10.全力推进“数智贵阳”（二期）项目建设。启动块数据（城市）综合服务平台、块数据（城市）运行调度中心、智慧市场监管等项目建设，有效提升政府治理能力。2020年力争完成投资3.5亿元。（牵头单位：市大数据局、市财政局、市直相关部门、市大数据集团）

11.做优做强“数智贵阳”平台入口功能。深入推进“一网通办”平台建设，接入1200项以上政务服务事项，在政务服务事项网上可办率100%的基础上，实现政务服务“零跑腿”全程网办事项占比70%以上。完成10个市级部门高频APP（公众号）接入“数智贵阳”平台工作，努力将“数智贵阳”平台打造成为企业和群众办事的总移动端。（牵头单位：市大数据局、市政务服务中心、市直相关部门、市大数据集团）

12.完善市级政务大数据项目管理。出台《贵阳市市级政务大数据应用项目管理办法实施细则（试行）》及配套文件，开展项目全过程绩效管理工作，全面提升政务大数据项目管理水平。（牵头单位：市大数据局、市财政局）

13.深化“数据铁笼”建设工作。完成市工业和信息化局、市公安局、市人力资源社会保障局、市住房城乡建设局、市综合执法行政局等5家市直单位“数据铁笼”验收工作，全面完成“数据铁笼”一期工作。支持部分单位继续深化应用，开展“数据铁笼”二期建设。（牵头单位：市大数据局、市发展改革委、市财政局、市工业和信息化局、市公安局、市人力资源社会保障局、市住房城乡建设局、市综合行政执法局、市直相关部门）

三、进一步推动数据资源汇聚及共享开放

14.深入实施“数聚贵阳”工程。推进公安部大数据备份中心、公安部刑专数据中心、国家市场监管总局企业信用信息公示系统等国家部委数据中心建设。推进大数据安全产业数据区、南方电网（贵州电网）数据中心等行业数据中心落地。推动中电西南云计算中心、贵州翔明数据中心等第三方商业数据中心健康发展，综合利用率超过30%。（牵头单位：市公安局、市市场监管局、市大数据局、相关区〔市、县〕政府、相关开发区管委会、市大数据集团、贵州电网公司贵阳供电局）

15.完善政府数据共享开放考核机制。按照《贵阳市政府数据共享开放考核暂行办法》，对市直各单位，各区（市、县、开发区）政府（管委会）开展数据共享开放绩效考核工作，引导各部门，各区（市、县、开发区）政府（管委会）加快推进政府数据共享开放及数据应用工作。（牵头单位：市大数据局、市直各单位、各区〔市、县〕政府、各开发区管委会、市大数据集团）

四、进一步强化大数据创新能力建设

16.加快区块链生态培育。以“1432”工程为抓手，加快实施《贵阳贵安区块链发展三年行动计划》，2020年，“享链”城市级区块链基础设施平台建设初见成效，引入培育10家区块链重点企业、形成10个行业应用解决方案，初步构建区块链标准体系及测评体系。发布《贵阳区块链发展和应用白

皮书2.0》。（牵头单位：市大数据局、市投资促进局、市直相关部门、各区〔市、县〕政府、各开发区管委会、市大数据集团、中电科大数据研究院有限公司）

17.加快推进5G等新型数字基础设施建设。深入开展新型基础设施建设，完成建设投资25亿元。重点开展5G建设，完成《贵阳市5G发展规划（2020-2025）》编制，新建5G基站3064个，实现贵阳市核心城区5G网络覆盖。围绕智慧医疗、智慧教育、工业互联网等领域开展5个融合创新应用场景建设。（牵头单位：市大数据局、市发展改革委、市直相关部门、各区〔市、县〕政府、市大数据集团、中国电信贵阳分公司、中国移动贵阳分公司、中国联通贵阳分公司、中国铁塔贵阳分公司、贵州广电网贵阳分公司）

18.加快推动人工智能领域发展。以贵阳旷视人工智能产业基地建设为依托，构建以人工智能开放创新平台为核心的人工智能生态，推动优必选、翰凯斯、科大讯飞等重点企业项目建设，以应用为导向，加速人工智能应用创新和产业融合发展。（牵头单位：市大数据局、南明区政府、清镇市政府、高新开发区管委会、贵州双龙航空港经济区管委会）

19.加快推进“数典”工程研究。探索构建“数典”体系架构，打造数据基础资源集智平台。2020年初步完成术语词表库，案例库与模型库构建，上线“数典”开放平台，出版首部《数典》词典。（牵头单位：市大数据局、贵阳创新驱动发展战略研究院、中电科大数据研究院有限公司）

20.加强大数据标准体系建设。围绕重点领域发展需求,指导支持企事业单位积极参与大数据领域标准研制，力争发布10项大数据领域技术标准，培育5家标准创新型企业，扎实推进国家技术标准创新基地（贵州大数据）贵阳区域分基地创建工作。（牵头单位：市大数据局、市市场监管局）

21.深化大数据研发创新能力。完成提升政府治理能力大数据应用技术国家工程实验室验收工作。深化贵阳信息技术研究院（中科院软件所贵阳分院）建设，加速向新型研发机构转型，为贵阳市提升大数据研发创新能力提供有力支撑。（牵头单位：市大数据局、市发展改革委、中电科大数据研究院有限公司）

五、进一步加快数博大道高标准建设

22.加快建成数博大道核心区。启动数博大道大数据创新、产业聚集发展、城市综合管理等十大重点工程示范项目建设，确保全年建成10个以上数字化标杆项目。力争建成数博大道核心区，主营业务收入达1000亿元。（牵头单位：市大数据局、市直相关部门、云岩区政府、南明区政府、白云区政府、观山湖区政府、高新开发区管委会、市大数据集团）

23.加强数博大道产业发展统筹。以数博大道延伸段建设为契机，做好贵阳市贵安新区产业发展统筹。建立数博大道贵阳市贵安新区协同工作机制，制定数博大道延伸段方案，共同打造数博大道产业园。（牵头单位：市自然资源和规划局、市交委、市大数据局、市直相关部门、南明区政府、花溪区政府）

贵阳市大数据发展工作领导小组办公室 2020年4月13日印发

/ 工信部：运用大数据5G等技术 助力中小企业复工复产 /

来源 / 工业和信息化部　编辑 / 协会会员处 李苗苗　日期 / 2020-04

工业和信息化部办公厅关于开展2020年中小企业公共服务体系助力复工复产重点服务活动的通知

各省、自治区、直辖市及计划单列市、新疆生产建设兵团中小企业主管部门，有关单位：

为深入贯彻落实党中央、国务院关于统筹推进新冠肺炎疫情防控和经济社会发展工作决策部署，推动中小企业健康发展，现就开展中小企业公共服务体系助力复工复产重点服务活动通知如下：

一、总体要求

全面加强中小企业公共服务体系建设，紧紧围绕中小企业复工复产和高质量发展开展重点服务活动，解难点、除痛点、疏堵点、补盲点，为中小企业恢复生产经营和可持续发展切实提供支撑和保障。

二、重点服务活动

（一）政策宣贯服务。通过开设网上政策服务专栏、编发政策指引等方式，广泛宣传国家和地方出台的系列惠企政策。重点宣讲解读直接关系中小企业权益的财税支持、金融支持、社保减免、劳动用工等政策，汇集发布申报渠道和流程，帮助企业用好用足政策，打通政策落地“最后一公里”。

（二）数字化赋能服务。推动实施《中小企业数字化赋能专项行动》。聚焦线上办公、远程协作等方面，引导数字化服务商提供解决方案、工具包、工业APP等数字化服务产品。

强化智能制造服务，帮助企业加快数字化改造，支持中小企业设备上云和业务系统向云端迁移。举办“创新中国行”数字化应用推广等活动。

（三）创业创新服务。开展研发成果转化等创业服务，举办技术难题揭榜、诊断咨询等活动，优化创业创新环境。举办大中小企业融通对接、双创示范基地“融通创新”主题日等活动，推广“龙头+孵化”等融通发展模式。组织企业参加“创客中国”中小企业创新创业大赛和全国“双创”活动周，推动项目落地和投融资对接。搭建产业链供需对接平台，开展生产要素供需对接服务，助力产业链固链、补链、强链。

（四）“专精特新”企业培育服务。建立完善“专精特新”中小企业培育库，为入库企业提供技术创新支持、知识产权托管维权、品牌宣传推广等专项服务，促进其成长为专精特新“小巨人”企业、制造业单项冠军企业。开展“专精特新——腾计划”等活动，助力企业借助电子商务升级转型。

（五）融资服务。推动金融惠企政策落实，梳理摸排中小企业融资需求，加强与金融机构联系合作，推动其为中小企业提供信用贷款以及应收账款、订单、仓单和存货质押融资等金融服务。发挥政府性担保、再担保机构融资增信分险作用，助力中小企业复工复产。开展优质中小企业上市培育，促进投融资服务对接，提高中小企业直接融资比重。

（六）市场开拓服务。搭建线上产销对接平台，组织企业开展网上洽谈、在线签约等灵活多样的营销和招商活动。指导企业建立网上直播间、网上会客厅、新媒体营销平台，构建企业与电商平台对接桥梁，助力企业快速拓展销售渠道。支持企业运用招标采购平台和中小企业自采平台，实现网络化招标采购。

（七）其他专业化服务。举办中小企业线上人才招聘等活动，助力补足复工复产用工缺口。开展“企业微课”等线上培训活动，邀请知名专家、企业家在线授课，提升中小企业经营管理水平。加强法律援助和法律咨询公益服务，帮助企业解决受疫情影响造成的合同履行、劳资关系等法律问题。开展志愿服务，建立专家志愿服务团，充分调动社会力量服务中小企业。

三、保障措施

（一）加强组织领导。各地中小企业主管部门要加强对服务体系建设的组织领导，加大资金支持力度，结合本地实际制定具体实施方案，细化工作措施，因地制宜开展特色服务活动，全面助力中小企业复工复产。

（二）提升服务能力。强化中小企业公共服务平台网络、中小企业公共服务示范平台、小型微型企业创业创新示范基地和创新创业特色载体的带动作用，推动服务机构加强能力建设，促进资源共享和服务协同，完善评价机制，提高服务质量。发挥全国中小企业服务联盟作用，通过举办能力竞赛等活动，推动提升服务实效。

（三）创新服务方式。充分运用大数据、云计算、人工智能、5G等新一代信息技术，创新服务方式、拓宽服务渠道，有针对性地推出复工复产服务包、租金减免优惠包等专项服务产品，通过“互联网+”服务等形式，精准满足中小企业需求。

（四）及时总结宣传。各地中小企业主管部门要注重梳理总结经验做法和典型案例，加强分享借鉴和宣传推广，并及时将相关信息发送至cxfwc@miit.gov.cn。请填写重点服务活动实施情况统计表，形成重点服务活动工作总结，于年底前报送部（中小企业局）。

/ 阿里巴巴数据分析实战：超详细的母婴电商分析流程 /

来源 / 工业和信息化部 编辑 / 协会会员处 李苗苗 日期 / 2020-04

随着科技互联网的发展，电子商务在现代商务企业的发展中占有越来越重要的地位，而数据分析作为电商行业非常重要的一种运营手段，在营销管理、客户管理等环节都需要应用到数据分析的结果。本文以阿里巴巴天池的婴儿用品购买数据为例，进行相关分析并提出建议。

数据理解

1. 数据来源

笔者在阿里巴巴天池下载的数据，这部分帮大家整理好了，可以在文末添加客服领取。

2. 字段含义

表名	字段	含义	备注
表1 购买商品	user_id	用户ID	购买账号
	auction_id	物品编号	
	cat_id	商品种类ID	商品二级分类
	cat1	商品种类ID	商品一级分类
	property	商品属性	描述商品特征
	buy_mount	购买数量	单笔购买数量
	day	购买时间	购买日期
表2 婴儿信息	user_id	用户ID	购买账号
	birthday	生日	婴儿出生时间
	gender	性别	0女性；1男性；2未知的性别

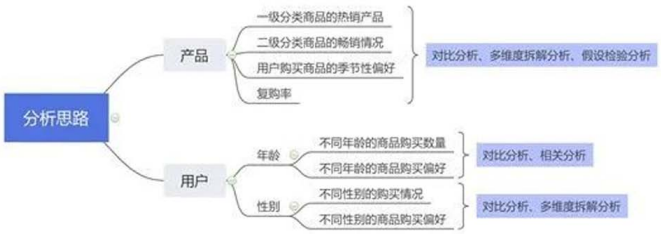
分析思路

1. 问题

- 年龄段的分布情况
- 哪一类商品最受欢迎
- 不同性别对商品的选择情况
- 同一商品大类下不同分类的畅销程度
- 用户购买商品的季节性偏好

2. 思路

从产品、用户两个维度展开：



数据清洗

1.在此次的数据分析中不需要用到表1的“物品编号”、“商品属性”列，隐藏该列。

2.修改列名，改成中文。

3.表1为用户购买记录，不能排除同一位用户在同一天重复购买相同商品的情况，所以不做重复值剔除，表2用户ID无重复值。

4.表1各列计数均为29972，表2各列计数均为954，未发现缺失值。

5.生日、购买日期通过分列功能转化成日期格式，性别通过if函数转换成文本“女、男、未知”，通过vlookup函数多表关联查询功能及datedif函数求得购买年龄。

6.通过排序和筛选功能，将购买年龄为28岁的数据进行异常值剔除。

用户ID	商品种类ID（二级）	商品种类ID（一级）	购买数量	购买时间	生日	购买年龄	性别
513441334	50010557	50008168	1	2012/12/12	2011/1/5	1	男
377550424	50015841	28	1	2012/11/23	2011/6/20	1	男
47342027	50013636	50008168	1	2012/9/11	2010/10/8	1	男
119784861	50140021	50008168	1	2012/11/29	2012/3/27	0	女
159129426	50013711	50008168	2	2012/8/8	2010/8/25	1	女
645596397	50010549	50008168	1	2012/11/30	2013/2/20	未出生	女
757254614	50013711	50008168	1	2013/12/20	2009/5/28	4	女
275261625	50010558	50008168	1	2013/6/10	2010/5/25	3	女

问题分析

1. 产品维度

（1）一级商品种类按季度汇总销量如下

求和项:购买数量	一级商品种类				
时间	28	38	50008168	50014815	5002252
2012年	2357	466	2029	1446	400
7月	237	65	175	111	38
8月	237	121	215	131	64
9月	394	59	558	233	77
10月	383	65	335	181	67
11月	855	86	405	577	87
12月	251	70	341	213	67
2013年	11217	1119	5209	3255	1267
1月	638	54	338	212	86
2月	827	78	100	125	34
3月	261	63	324	270	85
4月	612	68	387	289	86
5月	547	91	631	252	263
6月	412	109	289	222	160
7月	1559	98	614	289	82
8月	535	119	355	239	77
9月	904	125	532	242	87
10月	562	89	542	222	101
11月	1124	92	550	530	127
12月	3236	133	547	363	79
2014年	12411	1779	10830	14600	1460
1月	358	86	353	193	60
2月	722	147	202	247	104
3月	1195	87	553	292	127
4月	713	125	751	376	130
5月	1740	108	1076	505	124
6月	647	94	507	354	111
7月	1446	152	385	315	111
8月	988	236	1086	656	112
9月	1044	101	3500	292	127
10月	1704	175	799	367	174
11月	973	261	896	10586	150
12月	881	207	722	417	130
2015年	2560	302	723	462	118
1月	2378	159	616	411	97
2月	182	143	107	51	21
总计	28545	3666	18791	19763	3245

发现2012年只有三、四季度的销量数据，2015年只有一季度的销量数据且并不完整。为了保持数据完整性，故选取2013、2014年进行分析。

购买量	一类商品种类ID				
购买时间	28	50014815	50008168	38	50022520
2013	11217	3255	5209	1119	1267
2014	12411	14600	10830	1779	1460
总计	23628	17855	16039	2898	2727
增量占比	110.64%	448.54%	207.91%	158.98%	115.23%

从年度销量数据来看，28居销量第一，其次是50014815，但2014年50014815销量激增反超28夺得销量榜首，28增长率在品类中倒数第一。现通过假设检验分析方法进行具体分析。

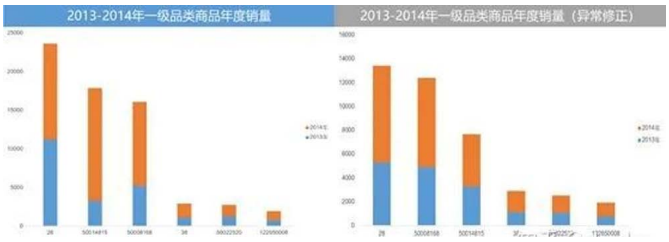
假设一：少数客户购买销量巨大，对各品类商品总销量有明显影响

对购买量进行描述性分析，如右图：

大多数的订单购买量为1，最大单笔购买量为10000，标准差为63.99，说明存在少量客户大量购买导致销量剧增的情况，剔除与平均值偏差超过3倍标准差的异常值，即购买数量大于192的数据样本。

购买数量	
平均	2.544125988
标准误差	0.369607104
中位数	1
众数	1
标准差	63.98687887
方差	4094.320667
峰度	20133.78257
偏度	133.4585459
区域	9999
最小值	1
最大值	10000
求和	76250
观测数	29971
最大(5)	1000
最小(5)	1
置信度(95.0%)	0.72444587

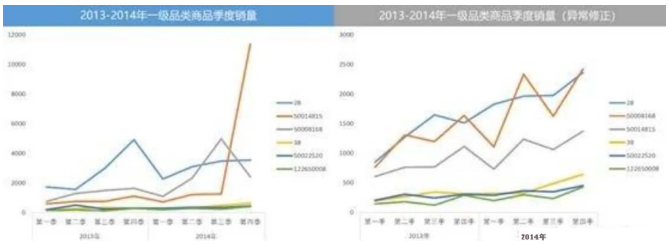
进行剔除前后年度销量数据比较如下：



由上图可以看出，2014年50008168的销量激增很大一部分原因来自个别客户的大量购买，剔除异常数据后，28依旧为用户最受欢迎的选择，假设一成立。

假设二：促销活动有带动作用

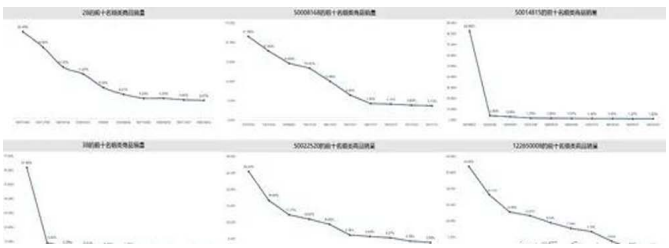
进行剔除前后季度销量数据比较如下：



由上图可以看出，122650008、50022520、38季度销量比较稳定，28、50008168和50014815销量最多主要在四季度，考虑到大客户订单对销量影响明显，提取这三个品类剔除异常值的四季度日销量进行分析。

可以看出28、50008168和50014815的销量峰值主要集中在11月11日和12月12日前后，推测为双十一和双十二促销活动带动销量的增长，假设二成立。

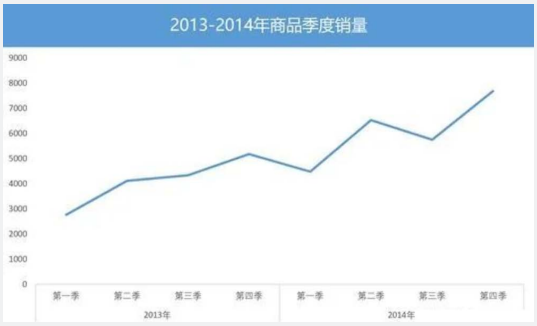
（2）对一级商品种类内销量前十的二级细分商品销量进行分析



由上图可知，50014815和38中二级商品细类销量排第一的分别是50018831、211122，都在60%以上，各商品销量差距大；而一级商品种类中28、50008168、50022520和

122650008购买量排名前2的二级细类商品占比均在40%左右，各商品销量较为均衡。

（3）为了保证数据的完整性和排除个别客户大量订单对数据造成影响，选取2013-2014年异常修正后的数据进行分析。



商品销量在二季度、四季度有增长峰值，对这两个季度进行月销量分析。

二季度和四季度月销量最多的分别是5月和11月，对这两个月份进行日销量和细化一级种类商品销量分析。

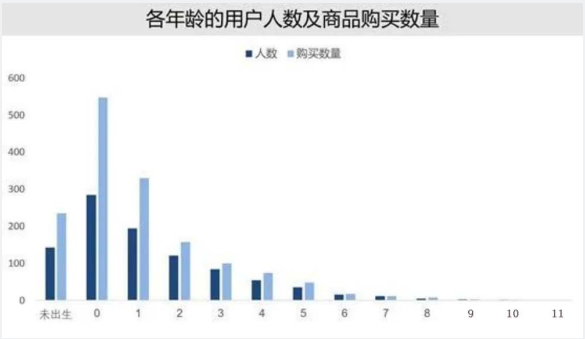
由此可以得出，5月销量集中在11日之后和月底，查询当年节日正值母亲节之际和儿童节前期，而11月份销量主要集中在11日前后，为双十一购物狂欢节时期。除此之外，无论是5月还是11月，销量贡献主要集中在一级品类商品28、50008168、38。

（4）复购率

总用户数为29943人，重复购买的用户数一共25人，复购率为0.083%，不到千分之一可以忽略，可能原因有①产品不被客户认可，缺乏用户年度②产品为耐用型。这部分需要额外的数据支撑，这里不再研究。

2. 用户维度

（1）年龄分布情况如下：

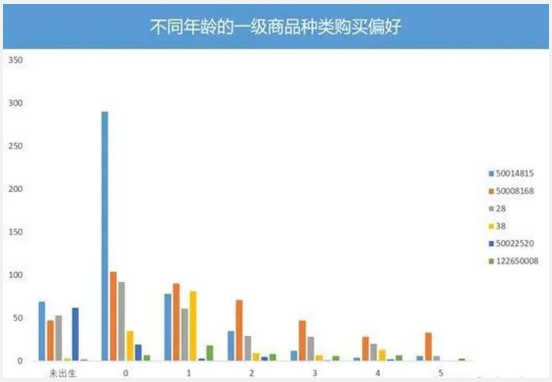


提出假设：婴儿年龄越小用户购买需求越大

进行相关分析:

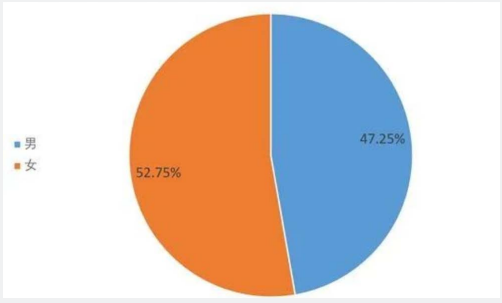
	年龄	购买人数
年龄	1	
购买人数	-0.85513	1

年龄与购买人数的相关系数为-0.86，高度负相关，证明假设成立。

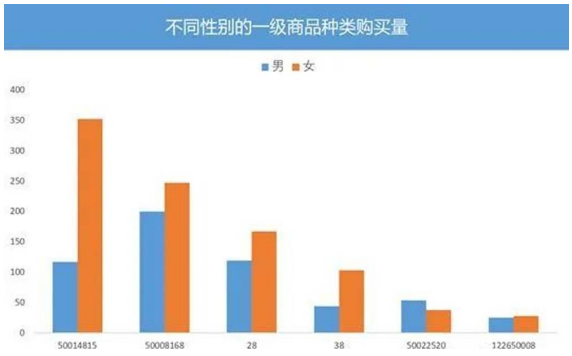


其中，各年龄层偏爱的一级商品种类前三分别是50014815、50008168、28。

（2）性别比例



在总销量中，女婴购买量占52.75%，相比男婴购买量的47.25%高出5.5%，说明女婴的购买需求大于男婴。



由上图可知，50014815最受女婴用户欢迎，50008168最受男婴用户欢迎。

结论建议

1.用户的购买量相差大，少数客户下单量巨大对销量有明显影响，建议做好大客户的服务，保持长期合作关系。

2.节日性活动对销量提升有明显促进作用，重点关注母亲节、儿童节、双十一和双十二。

3.销量贡献主要集中在一级品类商品28、50008168、38，根据商品的过往销量情况做好库存管理。

4.5岁以内是购买主力，随年龄增长购买需求下降，而女婴消费比例略高于男婴，后续策划促销活动可侧重借鉴此类用户画像。

5.复购率极低，要重点了解原因，如不是因为商品本身特性如耐用品，会有很高的提升空间。

/ 2020年中国新基建大数据中心产业链全景图深度分析 /

来源 / CPDA数据说　编辑 / 协会会员处 李苗苗　日期 / 2020-04

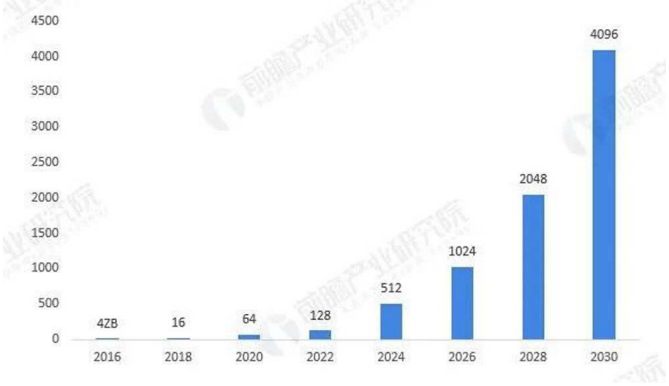
全球范围内，美国数据中心数量遥遥领先，亚太市场是全球数据中心市场的亮点，与2018年同期相比投资增长达到12.3%，增速遥遥领先。截止2019年中国数据中心数量大约有7.4万个，大约能占全球数据中心总量的23%，北京、上海、广州、深圳等一线城市数据中心资源最为集中。中国电信、中国移动、中国联通市场份额较大，具备资源优势。AI、5G、区块链等场景化应用以及工业计算需求助力行业发展，2025年中国数据中心IT投资将超7000亿元。

万物数据化另大数据中心的构建势在必行

互联网的高速发展使得万物数据化，数据量和计算量呈指数爆发，赛迪顾问数据显示，到2030年数据原生产业规模量占整体经济总量的15%，中国数据总量将超过4YB，占全球数据量30%。

数据资源已成为关键生产要素，更多的产业通过利用物联网、工业互联网、电商等结构或非结构化数据资源来提取有价值信息;而海量数据的处理与分析要求构建大数据中心。

图表1：2016-2030年中国数据规模增长预测(单位：ZB)



大数据中心定义

数据中心是指按照统一标准建设，为集中存放的具备计算能力、存储能力、信息交互能力的IT应用系统提供稳定、可靠运行环境的场所。数据中心按照服务的对象可以分为企业数

据中心(EDC)和互联网数据中心(IDC)。企业数据中心指由企业或机构构建并所有，服务于企业或机构自身业务的数据中心。互联网数据中心由IDC服务提供商所有，通过互联网向客户提供有偿信息服务。

按照机架数量规模可以将数据中心划分为以下五种类型。

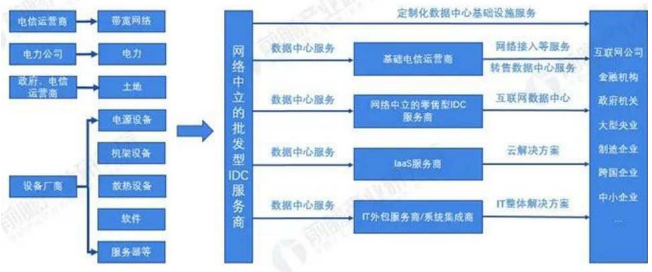
图表2：数据中心类型

数据中心规模	机架数量界定
超大型数据中心	>10000
大型数据中心	3000～10000
中型数据中心	500～3000
小型数据中心	100～500
微型数据中心（机房）	<100

大数据中心产业链

大数据中心产业链包括：上游基础设施及硬件设备商、中游为运营服务及解决方案提供商、下游为数据流量用户，温控设备是底层设施的保障。大数据中心是新基建的能量，汇聚了所有行业的数据、存储和分析，其重要性可见一斑，而大数据中心背景下，IDC和服务器是枢纽，也是行业最先受益的重要领域。

图表3：大数据中心产业链

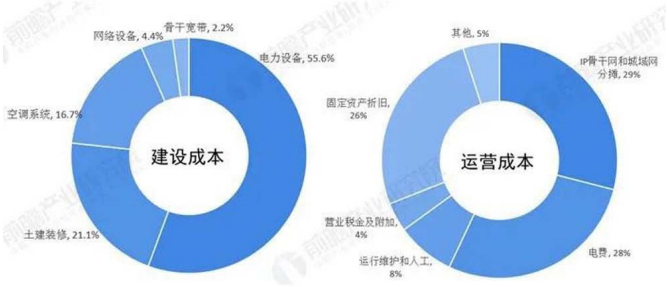


电力设备为成本大头，高压直流输电(HVDC)有望成为数据中心主流选择

低成本是数据中心运营商建立竞争优势的关键。数据中心成本由建设成本和运营成本构成。在数据中心的建设成本中，电力设备成本最高，其占比达55.6%。而与PUE指标关系密切的散热设备(为服务器、网络设备及电力设备提供空调散热)占比第三，为16.7%，核心设备一般5年进行更换替代。在IDc建设或改造中选用PUE更低的温控方案将成为国内IDc建设的必然选择。

而数据中心的运营成本的主要是IP骨干网和城域网分摊及电费，其占比分别为29%、28%。因此无论是供电设备还是电费，两者均是数据中心的成本大头。因此，降低电力基础设施采购成本，提高电源使用效率，是数据中心降成本的两大关键手段。

图表4：数据中心建设及运营成本分析(单位：%)

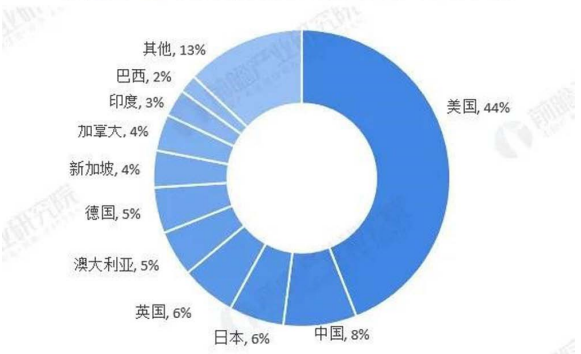


全球数据中心产业现状

1、美国数据中心数量遥遥领先

从全球范围来看，美国数据中心的数量最多，其占比达44%。其次是中国、日本、英国、澳大利亚、德国，其占比分别为8%、6%、6%、5%。

图表5：2018年全球大规模数据中心区域分布情况(单位：%)



2、数据中心投资稳步增长

云计算、大数据、物联网、人工智能等新一代信息技术快速发展，数据呈现爆炸式增长，数据中心建设成为大势所趋。世界主要国家和企业纷纷开启数字化转型之路，在这一热潮推动下，全球数据中心IT投资呈现快速增长趋势。全球及中国数据中心IT投资规模增长率均高于全球GDP增长率(2.3%)和中国GDP增长率(6.1%)。

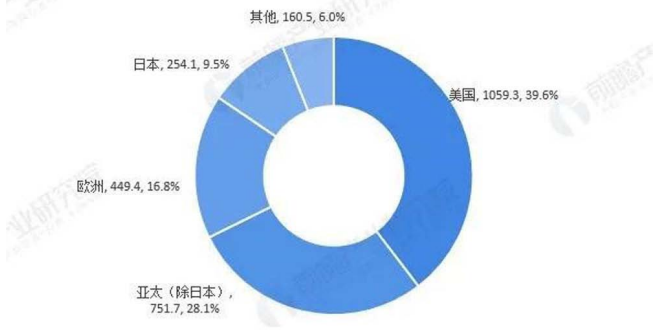
图表6：2017-2019年全球及中国数据中心IT投资市场规模及增长(单位：亿美元，亿元，%)

3、亚太市场投资增速最快

从全球数据中心建设发展来看，世界前三大数据中心市

场——美国、日本和欧洲的数据中心IT投资规模仍占全球数据中心IT投资规模的60%以上，美国保持市场领导者地位，在数据中心产品、技术、标准等方面引领全球数据中心市场发展。亚太市场仍是全球数据中心市场的亮点，与2018年同期相比增长达到12.3%，数据中心IT投资规模达到751.7亿美元，主要动力仍来自中国数据中心市场稳步发展，移动互联网、云计算、大数据、人工智能等应用深化，互联网+，人工智能x，工业互联网建设加速。

图表7：全球主要国家和地区数据中心IT投资市场规模及占比(单位：亿美元，%)



中国数据中心建设现状

1、数据中心机架规模

2019年中国数据中心数量大约有7.4万个，大约能占全球数据中心总量的23%，数据中心机架规模达到227万架，在用IDC数据中心数量2213个。数据中心大型化、规模化趋势仍在延续，区域性应用、多层级集团企业均倾向通过规模化建设避免盲目建设和重复投资。

图表8：2016-2019年中国数据中心机架规模变化(单位：万架)

2、IDC市场规模

中国IDC圈研究中心发布的《2019-2020年中国IDC产业发展研究报告》显示，中国IDC业务市场规模在日益增长的客户需求带动下仍保持稳定增长。2019年，中国IDC业务市场规模达到1562.5亿元，同比增长27.2%，增速放缓2.6个百分点，市场规模绝对值相比2018年增长超过300亿元。

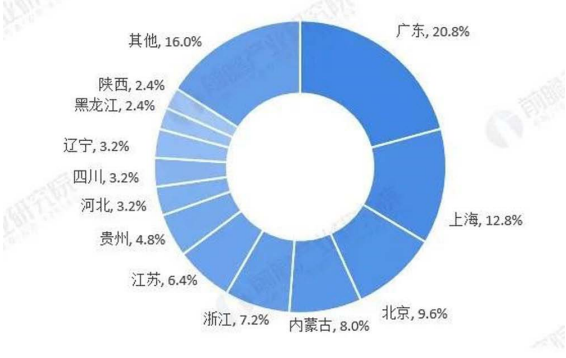
得益于5G、工业互联网以及人工智能等新技术的应用，各级政府部门，企事业单位纷纷加强了数据中心的建设及网络资源业务整合力度。在很大程度上推动了中国IDC行业客户需求的充分释放，拉升了IDC业务市场规模的持续增长。预计2019-2022年，中国IDC业务市场规模复合增长率将达到26.9%。

图表9：2014-2019年中国IDC业务市场规模(单位：亿元，%)

3、数据中心分布情况

我国对数据中心的需求主要集中在北京、广东、上海、浙江、江苏等经济发达省份，这些地区人口以及互联网用户密度远远领先中西部地区，互联网用户密度最大，大型互联网、云计算、科技创新类企业、政企用户数远远领先其他地区，对数字经济的贡献也更大，因此是我国数据中心业务需求最旺盛的区域。从全国范围来看，北京、上海、广州、深圳等一线城市数据中心资源最为集中，其上架率达到60%-70%。

图表10：2018年中国数据中心分布情况(单位：%)



此外，中部、西部、东北地区可用数据中心资源丰富，规模较大，价格优势明显。这些地区土地资源丰富，建设租金成本较低，网络质量、建设等级及运营维护水平也较高，适合建立大型及超大型数据中心。例如：百度最大的数据中心位于山西省阳泉市，服务器设计装机规模超过16万台;阿里江苏云

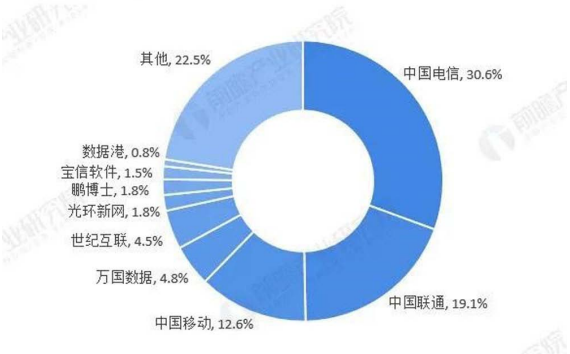
计算数据中心在南通签约，建成后将成为阿里巴巴华东地区最大的云计算中心基地，承载30万台服务器。

中国数据中心市场竞争格局

中国电信、中国移动、中国联通市场份额较大，具备资源优势。运营商核心优势在于对带宽资源的垄断，包括拥有大量机房、骨干网络宽带和国际互联网出口宽带资源。就市场规模而言，基础电信运营商占据着中国 IDC 市场约 65%的份额，但是一半以 上为自用，其他的机房遍布全国，在核心城市的 IDC 资源布局不多且客户较为分散。

且目前 的劣势在于 IDC 非主业，专业性不足，市场响应慢，局部供需不平衡，不符合市场微观需要， 且只提供各自网络接口，无法满足服务高时效和客户定制化需求。

图表11：2019年中国数据中心市场竞争格局(单位：%)



数据中心大型化、规模化成为行业发展趋势

2019年，超大型、大型数据中心数量占比达到12.7%，规划在建数据中心320个，超大型、大型数据中心数量占比达到36.1%。这一数据与美国相比仍有较大差距，美国超大型数据中心已占有到全球总量的40%，大型数据中心仍有较大的发展空间。

行业发展驱动性因素

1、AI、5G、区块链等场景化应用，为数据中心发展打开新的成长空间

在国家政策和资本的共同推动下，AI生态不断完善， AI场景化应用加速落地，AI基础设施服务服务将迎来快速发展新时期。5G商用在即，大量基于5G的应用在金融、制造、医疗、零售等传统行业中开始示范与推广，VR/AR、自动驾驶、高清视频、智能交通、智能医疗等应用需求也将为数据中心市场发展与服务模式创新打开成长空间。

2019年底，政府首次将区块链技术发展列为国家战略重点方向，未来，区块链技术在应用场景上将从当前的跨境交易、商品溯源、金融创新、供应链整合等经济领域，延伸到民生需求、城市治理和政务服务等社会政策和公共服务领域，必然带来大量分布式计算、分布式存储、分布式数据库管理需求，这些均离不开大数据中心做支撑。

2、工业计算需求旺盛，成为未来数据中心发展新动力

作为新一代信息技术与制造业深度融合的产物，工业互联网日益成为新工业革命的关键支撑，我国拥有巨大的信息化、数字化、智能化应用市场，传统行业的信息化改造、数字化及智能化升级，未来企业上云、设备上云步伐将加快，工业计算需求将爆炸式增长。

并且，国家高度重视工业互联网的发展，2019 年3 月工业互联网首度写入政府工作报告，提出要围绕推动制造业高质量发展，打造工业互联网平台，拓展“智能+”，为制造业转型升级赋能。工业互联网的应用部署与发展，工业计算服务需求将成为未来数据中心发展新动力。

2025年中国数据中心IT投资将超7000亿元

大数据中心背后有几个名词：数据库、区块链、数字货币、IT设备、云计算等。数字货币是区块链的应用之一，区块链是大数据的技术架构、云计算是区块链技术的技术支撑，IT设备是大数据的外在的储存之地。在海量数据存储的背景下，IT设备将迎来投资热潮，预计到2025年IT设备投资将达7070.9亿元。

图表12：2019-2025年中国数据中心IT投资规模增长预测(单位：亿元，%) 以下附大数据中心产业链企业名单：

图表13：大数据中心建设相关企业名单

数据中心设计	
北京电信规划设计院有限公司	上海邮电设计院
广东省电信规划设计院有限公司	同信通信股份有限公司
华信设计院	中国建筑技术集团
江苏省邮电规划设计院	
数据中心建设单位	
北京科海致能科技有限公司	上海华东电脑股份有限公司
北京长城电子工程技术有限公司	上海蓝色帛缔智能工程有限公司
浩德科技股份有限公司	上海窒峰计算机机房工程有限公司
捷通智慧科技股份有限公司	同方股份有限公司
青岛恒华机房设备工程有限公司	中国建筑技术集团有限公司

数据中心电气设备企业（UPS类）	
华为	美克电源
科华恒盛	维谛技术
康舒科技	先控
施耐德	伊顿
ABB	易事特EAST
冠军	英威腾
科士达	中达电通
数据中心电气设备企业（柴发类）	
德国MTU	三菱
道康	潍柴
科勒	威尔逊
康明斯	辛普森
科泰	玉柴
卡特彼勒	
数据中心电气设备企业（PDU类）	
华为	康普
施耐德	隆兴
ABB	南盾
昌遂	胜威南方
大唐	天乐
公牛	突破
华脉	图腾
金盾	天邮
克莱沃	维谛技术
数据中心暖通设备企业（空调类）	
华为	克莱门特
施耐德	美的
艾特网能	麦克维尔
大金	世图兹
格力	维谛技术
海洛斯	约克
海悟	依米康
佳力图	英维克
开利	
数据中心暖通设备企业（液冷类）	
广东合一新材料研究院有限公司	兆氟科技有限公司
深圳绿色云图科技有限公司	
数据中心运维企业	
北京创意银河电子科技有限公司	捷通智慧科技股份有限公司
北京健伦机房工程有限公司	青岛恒华机房设备工程有限公司
北京长城电子工程技术有限公司	上海华东电脑股份有限公司
北京真视通科技股份有限公司	上海金曜电子工程有限公司
福建长城电子技术工程有限公司	西安东升科技有限公司
数据中心运营企业	
北京云泰数通互联网科技有限公司	方正宽带网络服务有限公司
数据港	国富瑞数据科技有限公司
万国数据	光环新网
中昌大数据股份有限公司	华数云
中国电信	华通云
中国移动	鹏博士
北京多联元信息技术有限公司	秦淮数据
北京歌华有线电视网络股份有限公司	旗云科技有限公司
北京供销大数据集团	润泽科技
北京国研网络数据科技有限公司	世纪互联
北京蓝汛	首信融合
北京星缘新动力科技有限公司	云联数据
北京易华录信息技术股份有限公司	中国联通
东方有线	中经云
富春云	

数据中心IT设备企业	
安迈信息科技（昆山）有限公司（AMI）	联想
Cavium公司（CAVM）	曙光信息产业股份有限公司
东芝股份有限公司	赛灵思（Xilinx）
华为	神云科技
IBM	Western Digital
Intel	新华三集团
浪潮集团	中兴通讯
数据中心网络设备企业	
Amphenol ICC	赛尔网络有限公司
华工正源	上海欣诺通信技术股份有限公司
华为	盛科网络（苏州）有限公司
立讯精密工业股份有限公司	苏州旭创科技
擎发通讯科技（合肥）有限公司	武汉光迅科技股份有限公司
是德科技（keysight）	易飞扬(Gigalight)
索尔思光电	
数据中心用户企业	
阿里巴巴	京东
百度	美团云
国家电网	赛尔网络
华大基因	腾讯
华为	网易

图表 14：大数据中心企业排名

排行	IT基础设施企业	数据采集企业	数据整合分析企业	数据安全企业	服务器企业	IDC企业	存储企业
1	华为	阿里巴巴	用友	三六零	浪潮信息	中国电信	华为
2	中兴通讯	腾讯	科大讯飞	深信服	紫光股份	中国联通	紫光股份
3	用友	华为	浪潮信息	启明星辰	华为	中国移动	浪潮信息
4	深信服	字节跳动	广联达	卫士通	新华三	万国数据	联想
5	浪潮信息	美团	金蝶	绿盟科技	广达电脑	有孚网络	中科曙光
6	中科曙光	京东	美亚柏科	美亚柏科	联想	世纪互联	光环新网
7	中际旭创	滴滴	东方国信	安恒信息	中科曙光	光环新网	深科技
8	金蝶	海康威视	华胜天成	迪普科技	华硕	宝信软件	易华录
9	清华同方	网易	华数传媒	东方通	航天联志	鹏博士	华胜天成
10	新华三	百度	数知科技	北信源	信维股份	数据港	东方国信
11	联想	小米	数据港	山石网科	长城超云	科华恒盛	数据港
12	星网锐捷	中兴通讯	捷成股份	蓝盾股份	清华同方	网宿科技	紫荆储存
13	网宿科技	科大讯飞	拓尔思	任子行	宝德股份	浙大网新	捷成股份
14	东软	恒生电子	久其软件	天喻信息	亿时空	奥飞数据	科华恒盛
15	新易盛	启明星辰	海量数据	安博通	杰和科技	首都在线	同有科技
16	中软国际	四维图新					海量数据
17	英维克	东方国信					英科科技
18	科士达	神州信息					朗科科技
19	科华恒盛	宜通世界					天玑科技
20	银信科技	易联众					宝德股份

/ 数据分析师最容易跳进去的5个大坑！原来就在我们身边！ /

来源 / CPDA数据说 编辑 / 协会会员处 李苗苗 日期 / 2020-05

作为数据分析人员，每天提取数据、写分析报告，如此反复却总得不到提高？又或者经常迷茫不知道自己的价值在哪里？花了很久做了一份详细的报告却没有什么用？看完本文或者会有些领悟，有的时候换一种方式思考，效果会更好~

移动应用和手机游戏的飞速发展催生了巨量级的数据资料，这些数据生动地刻画了用户的使用轨迹和行为习惯，价值难以估量。

于是，针对这些数据开展专业研究工作的数据分析人员成了香饽饽，他们的分析结论有可能对一个产品的发展走向带来巨大的影响。而作为数据分析人员，要在大量的数据中找到有意义、有价值的内容并不是易事。

过去，数据分析师绝大多数来自统计学或编程学的人才。随着越来越多企业发现，数据分析人员应该同时具备数据分析能力以及商业运作能力，这种情况在近几年才有所转变。对数据的解读能力、问“正确”问题的能力以及解答问题时的灵活性，都是衡量一名数据分析人员是否足够称职的关键。

数据分析师Pavel Trejbal持有认知信息学硕士学位，就职于AppAgent（为移动游戏工作室和创业团队提供营销服务的一家企业）。他的学术领域涉及到许多学科，包括经济教育学、心理学、脑科学、语言学、人工智能以及哲学。Pavel表示：“我不敢妄言自己是这些领域的专家，不过对这些领域的广泛认识的确帮助我在面对难题时以出其不意的角度找到解决方法。”

在数字的海洋里翻滚了六年，Pavel有过不少成功的表现，也有过很糟糕的分析结论。在这里，他给我们分享了数据分析人员最常犯的5个错误，以及对应的预防方法/建议。

错误1：执着于完美的算法

明明有现成的、简单的但非常适用的方案不采用，偏偏把时间花在对数据算法的钻牛角尖上，这是数据分析人员所犯的最常见的错误。与其花上一整个月的时间交出一份无比详尽的长文报告，不如在短时间内交出一份简洁的数据分析。也许后者在一些细枝末节上不够精确，但具有直接参考价值的结论才是你的上级亟需的。直击要点才是最有效率的做法，在商业战争里时间太重要了！

错误2：迷信完美通用的方法论

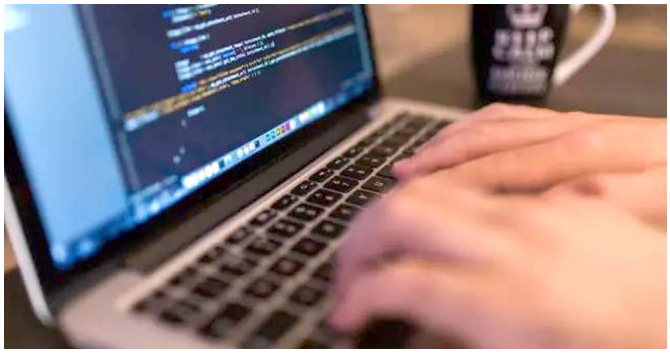
千万不要这样做。每一个业务，每一次分析，都是有区别的。完美通用的方法论听上去很美好，但具体的方案必须由自己思考得出。对待每次分析，都应该是面对全新挑战的姿态，开放思考、亲自分析，不能依赖过往的类似案例。

错误3：只看数据，忽视其他分析依据

如果在数据分析过程当中发现一些特别突出的数据变化，记住：三人行，必有我师焉。在定论出来之前，主动找到产品运营、社区运营或者游戏策划商量，毕竟这些同事才是与用户有最直接接触、最理解产品的人。异样的数据变化，经常来自于不科学的解读方法或者数据采集过程中的技术错误。

错误4：清理数据的方式不科学

清理数据在数据分析工作里是个比较无趣的工序，而且往往要花上大部分的时间，但这个工序是绝对不能忽视的。在清理数据的过程当中，你会了解到哪些地方分析错了或者遗漏了、哪些地方限制了你的解读能力。如果跳过这个工序，分析结果很可能不靠谱，甚至得出与客观情况完全相反的结论。



错误5：无法分辨不同的工具和指标

因为总会存在不同的技术设定或者指标定义，所以每一款数据分析工具都是独一无二的。使用这些工具之前，一定要清楚区别在哪里。

最近我们就有用Google Analytics采样分析里的转化率和收入数据来进行A/B测试。刚开始，A变量在两项指标中都比B变量有更好的表现，但我们没有直接采用这个结论。我们把原始数据下载下来进行手动的分析。这次的分析结果跟之前完全相反，A变量在两项指标中都比B变量差很多。

离开座位，多多走动

Pavel确信，身为数据分析人员，无论如何都不应该守在自己的“象牙塔”里。相反，数据分析人员应该更多参与到公司的日常业务里，比如出席运营营销体系、产品策划团队的会议。如此，数据分析人员才能更好地理解策划人员及决策者的需求，接收更多跟产品直接相关的信息，并且适时提出数据分析提高产品表现的方案。除此之外，决策者们也更能理解到数据分析的价值，并且激励整个团队的钻研精神。

数据分析是非常重要的一环，虽然很复杂，但掌握一定逻辑和方法后，应该说不会有太多难处。而且，这不有我们作为前车之鉴吗？请不要再犯这些错误才好。

/ “五一”旅游市场大数据：“后浪”已成主力军 /

来源 / 经济日报 编辑 / 协会会员处 李苗苗 日期 / 2020-05

疫情稳定、天气尚佳，多重利好下，今年“五一”假期，全国各地旅游消费的需求集中爆发。5月5日，国家文旅部发布数据显示，从5月1日-5日，全国共计接待国内游客1.15亿人次，实现国内旅游收入475.6亿元。

5月5日，携程综合大平台相关数据发布的《“五一”旅游市场复苏大数据报告》显示，虽然总数上还未恢复到去年同期水平，但这个五一假期注定“意义重大”，疫情使各地景区纷纷加速在线预约建设，在线订票人数不断增长；“后浪”成为主力出行人群；景区在预约制下舒适度大幅提升；各类新玩法层出不穷……

“从国内旅游业绩看得出：大家的信心在逐步恢复。我们要做的不是期待反弹，而是努力推出更适合消费者的产品，更好的优惠套餐，通过营销创新推给消费者，加速行业恢复过程。”携程联合创始人、董事局主席梁建章说。

5天总人次破亿

疫后首个旅行高峰

“五一”假期前，携程在《2020“五一”旅游消费新趋势大数据报告》中预测：今年“五一”小长假，国内旅游市场将迎来新冠疫情后真正意义上的第一个旅行“高峰”。

5月5日，国家文旅部发布总结数据，结果显示，5月5日当天，全国接待国内游客1023.1万人次，实现国内旅游收入43.3亿元。

长达5天的假期，成为旅游市场复苏助推器。携程总结数据显示，在4月29日预订4月30日-5月5日出行的交通订单（含机票、火车票、汽车票、用车），对比4月1日-4月23日期间预订4月24日-4月29日交通出行订单量，增长幅度超过130%。

其中，4月29日当天，预订4月30日-5月5日出行的机票订单量，对比4月1日-4月23日期间预订4月24日-4月29日的机票订单量，增长51%。

不止是机票。疫情影响下，自驾成为许多游客出行的最主要选择，带动省内游、短途周边游快速发展。统计显示，“五一”携程租车预订单已恢复至去年同期水平，并且还有10%左右的增长。其中，无接触的在线租车成为热门选择，目前，携程与供应商合作已可以提供20万台车、3000多款车型，覆盖境内700多个热门城市与目的地、5万多个取车门店。

“旅客们的旅行热情，之前一直被疫情压制。但其实早在2月，我们通过携程机票爆发性的搜索数据就发现，大家对‘五一’小长假就充满了期待。”携程相关负责人表示，“与此同时，近阶段仍是国内旅行价格最好的时间段，价格实惠，各平台也推出了涵盖众多旅行产品的优惠活动。”

“比如，截止到4月25日，上海-深圳机票价格同比降低13%，上海-成都同比降低29%，上海-重庆同比降低33%，上海-三亚同比降低19%，广州-上海同比降低35%，深圳-成都同比降低33%。”上述负责人介绍。

成都、三亚、上海受青睐

“后浪”成旅游主力军

从目的地看，携程总结报告显示，成都、三亚、上海等城市成为各地旅客最青睐的几个目的地之一。

“我们跟踪了三亚、上海、南京、广州、昆明、杭州、贵阳、西安等城市的旅行情况，这些城市的旅客爱去的目的地，均包含了成都、三亚、上海这三座城市。而这三座城市，近期也一直在力推相关的旅行产品。”携程相关负责人称。

以上海为例，近期为了迎接上海发起的“五五购物节”，携程将通过旅行直播新玩法、安心游联盟新保障、优惠券+产品补贴高性价比等各种模式，为上海消费者补贴10个亿，这些优惠措施将从5月初密集释放到6月，线上线下全覆盖。

“携程的优惠补贴包括1亿元优惠券，内容涉及10~100元不等的酒店、用车、门票等补贴；9亿元产品补贴，推出包括酒店6折起、预售5折起、门票5折起、亲子酒店8.5折起等各品类产品折扣，优惠总价值超过10亿元。”

携程CEO孙洁表示，今年“五一”期间，上海已有超过170家景区可以在携程上预约预订门票，部分热门景点如海昌海洋公园还有低至3.8折的优惠。同时，为了防止游客聚集，携程全力支持景区“在线预约制”，以及秒入园、无忧退、线下票机、语音导览、分时预约等服务，帮助游客进行无接触和预约购票。

此外，今年“五一”期间，自由随性的年轻人群成为旅行主力军。结合携程省内跟团游、自由行产品的订单情况看，90后、00后占比超过一半，超过70后等“前浪”。其中，90后、00后的人气目的地有九寨沟、成都、张家界、上海、苏州、三亚、北京、重庆、常州、荔波、长沙、安徽黄山等。

同时，“五一”出现多种国内旅游新方式，全新开发上线的一些产品受到“后浪”群体的青睐。据携程“主题游”平台统计显示，徒步登山、户外、房车、越野、露营、旅拍，在“五一”主题游产品中在线浏览预订人气领先。

预约制成主流

高品质、安心游受追捧

与往年“五一”小长假不一样，今年“五一”国家对景区提前发出了完善预约制度，接待量不超过30%最大承载量等要求，“不预约，不出游”成为今年五一最显著的特点之一。

报告显示，假期前，不少景区都提出了提前设置客流额度，超出不售票的要求，使大量游客提前通过携程等在线旅游平台订票。携程门票数据显示，超过4000家景区已可以在携程预约预订门票。假期前三天，通过携程预约景区门票的人数相比清明增长176%，相比去年“五一”同期，门票预约人数也恢复至50%左右。

预约制的流行，也使得今年“五一”成为有史以来最“智慧化”的小长假。全国各个景区纷纷开启智慧建设步伐，并与携程等在线旅游平台合作提出智慧景区服务标准。游客通过携程等平台，即可享受实名认证、无接触服务、秒入园、无忧退、语音导览等服务。

今年“五一”与往年另一个不同是，根据携程总结数据，今年的“爆款”普遍集中在“高品质”和“安心游”产品上。

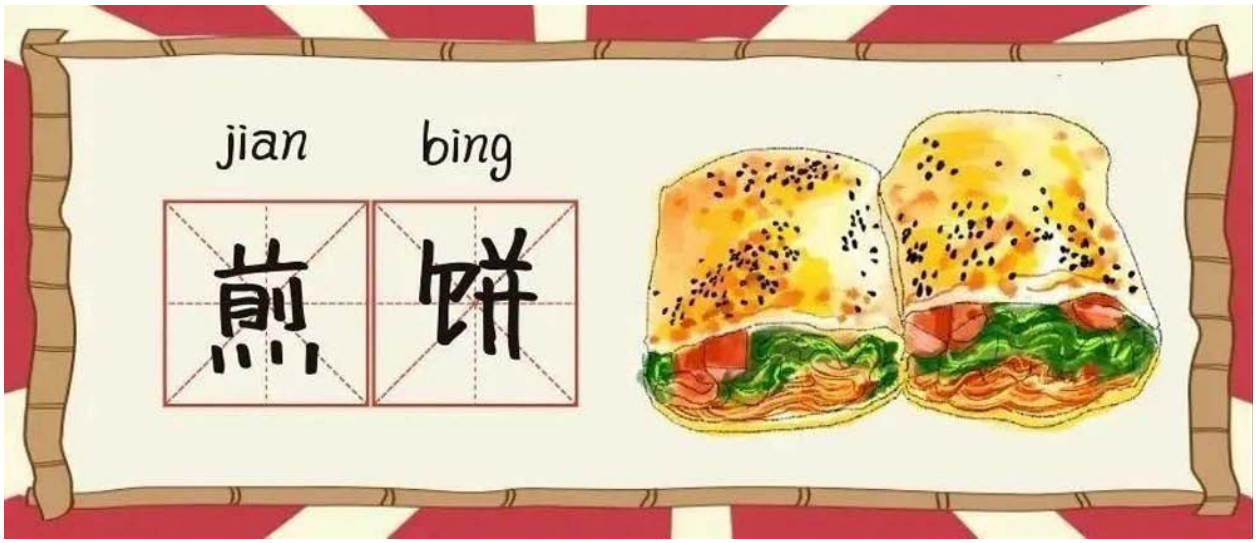
数据显示，“五一”期间省内自由行人气持续复苏，5钻高端产品占自由行订单总量的70%以上。省内“新跟团”产品在疫情背景下也相对更受欢迎，以携程主打的私家团为例，上线产品超过1500条，平均每团约2到6人，“五一”订单比清明小长假增长了两倍。

数据显示，“五一”游客出行，多选大面积、坐拥山景、湖景、海景、私家沙滩的高星酒店。其中，从房型选择上，相比往年，套房、别墅类房型销量略有上升，亲子房仍是五一游客刚需；从酒店星级上来看，消费者已预订的酒店中，4星、5星间夜占比达55%，高星级酒店明显更受欢迎；在携程的酒店预售中，5星级酒店的间夜占比达到了50%，选择1000元以上酒店产品的订单比例也在全部订单中占比最高。

同时，“安心游”为今年五一出游主旋律，也获得了旅客们的认同。“五一”前夕，携程发布了跟团“安心游联盟”和服务标准，上线6000多条“安心游”标签产品，从携程自营跟团产品的订单看，70%旅客选择了带“安心游”标签的产品。

/ 你信了摆摊经济，推着小车出门以后 /

来源 / CPDA数据说 编辑 / 协会会员处 李苗苗 日期 / 2020-05



互联网人，来摆摊了！摆摊经济一词大火以后，很多自媒体又开始算起摆摊的账。特别喜欢算出诸如“卖煎饼大妈月入3万”之类的账目，再加个“互联网人”的点缀，刺激在西二旗或者张江高科地铁站拼了老命才挤下车的互联网人的眼球。

这种文章权当调侃，看看就好了，可如果真的有人信了呢……

1、看似科学的分析过程

互联网公司的小王就真的信了！当然，他也按各种帖子的说法，做了细致的数据分析，励志要超越月入3万的卖煎饼大妈，成为煎饼王的男人！

在淘宝搜煎饼工具，获得固定成本：

· 全套煎饼工具：100元
· 小车：200元
· 炉子：200元

继续搜：生产成本：

· 鸡蛋5毛一个
· 面皮3毛一张
· 火腿5毛一条
· 大葱、酱、油成桶买的，预200元

这么算下来，一张煎饼成本1.5元，加火腿的也才2元，至少卖7块。每张毛利至少5.5元。月入30000，那么一天只要卖30000/（5.5*30）=180张就够了。一天早上班和晚下班能卖两拨，一拨只要90张。

当然，如此机智的小王，还想到了：

· 价值思维：在小车上带上矿泉水，纸巾，冰红茶，扩大卖点。
· 流量思维：一般企业上班早，6-8点卖一拨；广告公司、互联网公司的上班晚，9-10点换个地方卖一拨。
· 极值思维：煎饼一定要，摊的圆，蛋一定要涂抹均匀，打造核心爆款引流。

有互联网思维和大数据护体，这一波一定旗开得胜。之后还可以建立人工智能模型精准分析，稳赚稳赚！虽然早上出门的早点，晚上回家晚点，但是收入颇丰啊！再说了，搞得跟个我在互联网公司上班，不是早上6点出门晚上10点到家一样！于是满怀希望，小王凑齐小车出发咯！

2、真实上路的悲催

真上路以后：

· 第一天上路，遇到今年第一号热带风暴……
· 之后是，热带风暴二号、三号、四号、五号、六号……
· 风暴之后，是连续8天40度高温警报。突然觉得小车烤炉很多余啊，直接在马路上摊吧！

一个月后，以上终于消停了，能认真卖了吧！

1. 推到热闹的地铁站门口，吃到了当地摆摊的河南老乡们合力的痛打。损失小车*1
2. 推到没有团团伙伙的地方，发现摆摊的非常多，光煎饼摊就二十几个……
3. 推到既没有团团伙伙且没有人竞争的地方，发现没客人了……

跑遍大半个北京城，终于找到一个能摆摊的地方，这次能开搞了吧

1. 才发现摊煎饼原来这么耗体力，才摊了20张就累扑街了，感慨大妈身体真好；
2. 累的半死中午想吃顿好的补补，发现严重消弱毛利啊，大妈都是吃馒头喝凉水的……
3. 原来好地方没那么多，虽然城管不管了，但是写字楼、小区保安、实体店老板照样会怼；
4. 从一个区域推小车去另一个，浪费的时间比销售的时间多的多的多
5. 原来没几个顾客买煎饼又买其他的，推着一车东西还把自己累半死
6. 经常折腾销量不稳定，经常有多的面和蛋剩下还快坏了，干几天倒一批。好心疼……

到底哪里出了问题呢？

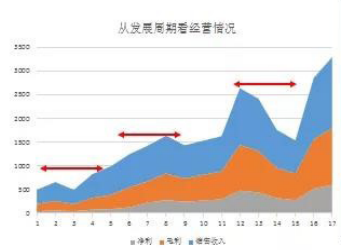
3、问题的核心在哪里

这里存在着3大非常常见的分析误区：

误区一：拿特例当常态。在产品经理、运营做案例分析的时候，这种事经常发生，往往缺少深入分析训练的人，喜欢拿最高光的时刻去做研究。从表现到本质，至少要稳定的观察一段时间，比如一到数个自然年，或者连续观察若干周期变化，这样才能沉淀经验。可做案例研究，往往喜欢抓住一个成功点猛吹，于是带来了误导（如下图）：

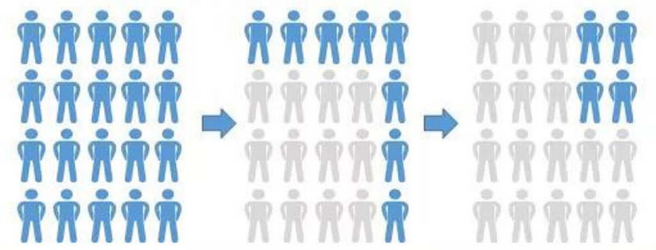
做数据分析的同学往往觉得做日报很低级，可恰恰是这些看似低级的日报，却能反应经营的真实形态。
类似煎饼摊大妈这种体力劳动者，受风吹雨淋影响明显，万一不出摊就没收入，因此**只有稳定观察一段时间，才能真正算清楚：到底营业额、毛利、净利是多少**

误区二：拿结果当过程。所有的胜利者分析都有这个通病。在胜利者胜利以后总结成功的十大因素，把胜利者无关痛痒的事迹，都算做成功的关键一发！可实际上，最后的结果都是一个增长过程里实现的，有可能在不同增长阶段需要关注的问题不一样，使用的方法、工具、甚至运气都不一样。（如下图）：



站在终点统计的数据往往会忽视起点的艰辛。比如左图，明显煎饼大妈历经了三个阶段，第一个阶段还不熟手，成本高，收入少；第二个阶段熟手了，收入慢慢提高，利润率也上去了，第三阶段可能换了摆摊位置，收入持续提高，利润率稳定。*《揆始气学盗防盗水印》***忽视起点的辛苦，往往会让新手栽倒在起步阶段**

误区三：拿个体当全体。很有可能这个最终胜利者只是一个幸运的幸存者而已，在观察留存情况的时候，整群看留存率更有意义，更能发现这到底是幸运儿还是能力真的强。



和我们做用户留存分析的时候，要群体看留存率一样，**脱离概率谈个案，很有可能谈的就是运气。**好的观察方法是：和大妈一条村出来摆摊的，有多少能坚持经营，多少赚到钱。*《揆始气学盗防盗》*同样做煎饼的大妈和卖水果的大妈，两个群体摊位换手率，经营月份哪个长，LTV收入哪个高

类似的事在生活中非常常见。我们总是听到一个成功以后的人或者企业在可劲吹，至于到底怎么发家的，到底从0到1从1到60的全貌是啥，除了他吹嘘的“拼搏”“努力”以外还有啥，通通不知道。可偏偏这种鸡汤最得人喜欢，因为听故事的有意思程度，远远超过搜集、整理、计算枯燥的数据。

有意思的是，我们会发现：老乡帮老乡的经营模式，能完美的避开以上三个坑。

1. 老乡们都在一起生活，了解所有业务常态。
2. 老乡们相互知根知底，能知道发展全过程。
3. 老乡们一条村干一个产业，所以容易观察到群体结果。

所以在真实的商业社会里，我们看到更多的成功例子都是一条村一条村的老乡抱团在做，潮汕商会、福建商会、做家具的江西人、以及摆摊最多的河南老乡。

朴素的商业逻辑后往往藏着最实在的道理，不是所有的经营模式运用“互联网思维”+“大数据”+“底层逻辑”+“人工智能”护体，就刀枪不入了，就地摊经济这种“小商贩”的经营模式而言，我们不仅要考虑从感性决策到理性决策的过程，更需要考虑从理性决策到行为决策的过程，否则那真是上边说的“胜利者鸡汤喝太多了”。

淘宝电商数据分析：1套真实+完整的案例分析流程

来源 / CPDA数据说　编辑 / 协会会员处 李苗苗　日期 / 2020-05

该数据分析借鉴的的背景数据来源于天池数据集，为2012年7月2日至2015年2月5日发生在淘宝天猫交易平台关于婴幼儿商品的交易数据。其中包括两个表格，截图如下：

	A	B	C	D	E	F	G
1	user_id	auction_id	cat_id	cat1	property	buy_mount	day
2	786295544	41098319944	50014866	50022520	21458:867	2	20140919
3	532110457	17916191097	50011993	28	21458:113	1	20131011
4	249013725	21896936223	50012461	50014815	21458:309	1	20131011
5	917056007	12515996043	50018831	50014815	21458:158	2	20141023
6	444069173	20487688075	50013636	50008168	21458:309	1	20141103
7	152298847	41840167463	1.21E+08	50008168	21458:340	1	20141103
8	513441334	19909384116	50010557	50008168	25935:219	1	20121212

	A	B	C
1	user_id	birthday	gender
2	2757	20130311	1
3	415971	20121111	0
4	1372572	20120130	1
5	10339332	20110910	0
6	10642245	20130213	0
7	10923201	20110830	1
8	11768880	20120107	1
9	12519465	20130705	1

涵括的字段有用户ID，交易编号，商品种类ID，商品类别，购买数量，购买日期，以及用户人为提供的个人信息如婴儿出生日期以及性别。字段含义解读如下：

· 用户ID：以电商行业的购买数据为例，用户ID是电商平台识别该购买者的唯一信息。从用户ID可以得到其他信息包括注册信息，购物历史记录，购物喜好等。

· 购买行为编号：我理解为交易号，通过这个号码可以查询到购买的具体事物，数量，购买行为产生日期，购买者或者需求者的位置。

· 商品种类ID：该电商平台继而又把每个类别下的产品细分了种类，即被购买产品属于该电商平台规定的某个类别的某个种类下面，并赋予每个种类一个ID.

· 商品类别：该电商平台把产品分成了很多个种类。

· 商品属性：即产品的详细情况

· 购买数量和购买时间即是字面意思

· 出生日期：记录的是该用户注册时填写的自己或者需求者的婴儿的出生日期

· 性别：即字面意思

分析目的

该分析旨在通过以往的数据总结以前的销售表现，找出需要改变及改善的地方，针对性采取有效措施以达到提升营业额的效果。

提出问题

2015年的销售下跌

· 第三，四季度销量上涨的原因

· 性别对销量的影响

· 年龄与销量的关系

分析思路

分析问题

由于整个分析过程都涉及到销量，所以在开始分析之前首先对购买量进行数据清洗。筛选购买量这一列可以发现，数据区间跨度非常大，对其作描述性统计发现，购买量的平均值不到3，标准差为65左右。

剔除与平均值的偏差超过三倍标准差的高度异常值，即大于199.64的数值都需要剔除。因为商品的单位不可能为小数，所以实际应剔除大于199的值。

购买数量	
平均	2.596862
标准误差	0.389611
中位数	1
众数	1
标准差	65.68376
方差	4314.357
峰度	19120.36
偏度	130.1
区域	9999
最小值	1
最大值	10000
求和	73808
观测数	28422
最大(5)	1000
最小(5)	1
置信度(95.0	0.763655

1. 2015年的销量下跌

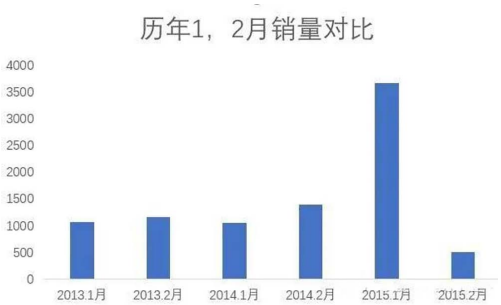
分析流程是这样的：

对购买量进行多维度拆解：购买量=新用户购买量+老用户购买量

新用户为首次出现，以前没有过购买行为的用户id，老用户为以前有过购买行为（重复的）的用户id。

通过查找重复值得知老用户为27个，占比为27/28396=0.93%，不到1%。换句话说，总购买量几乎全是由占比大于99%的新用户造成的，因此在这里我们忽略老用户的购买量。而且，从以往的销量折线图可以看出，2015年数据下跌是因为数据集里关于这一年的数据不全，只有1月和2月的数据。

假设在这一年里头两个月销量下跌，找出2015年的销售数据，同比历年的数据，来判断是否假设是对的。

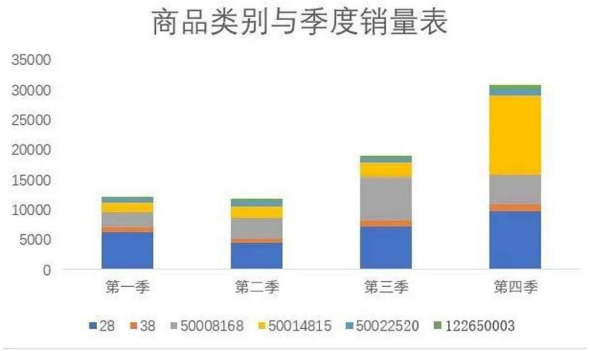


由于2012年缺乏上半年的数据，因此我们只能对比2013，2014和2015年销量。从图中可以看出，2015年1月销量大幅高于2013和2014年，2月销量低于前两年，但总和并不少于前两年。所以依据当前的数据不能证明2015年销量下跌，假设不成立。

2. 第三，四季度销量上涨的原因

分析流程如下：

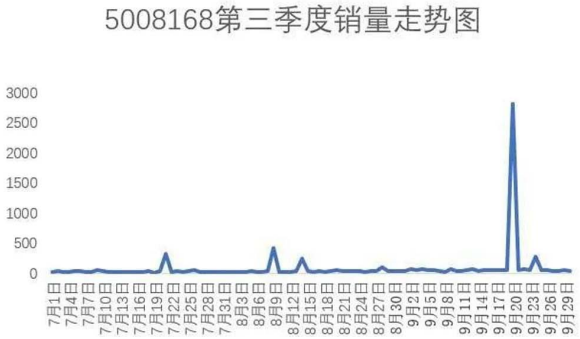
假设下半年销量上涨是因为所有类别销量上涨。我们提取各个季度各商品类别的销量数据，得到下图。



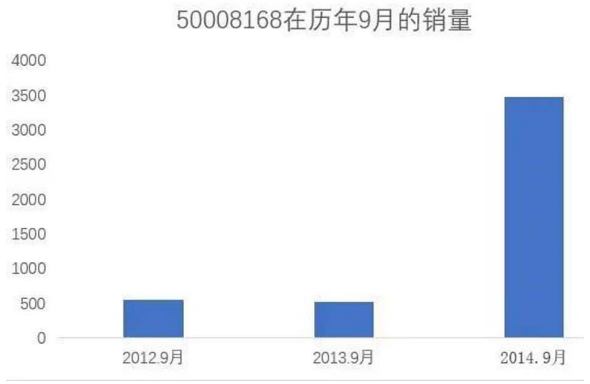
从上图可以看出，第一二季度销量基本持平，第三季度的销量主要是由类别5008168，和28带来的，其他类别没有明显变化。第四季度销量主要是由类别50014815，28带来的，其他类别差别不大。所以可以说季度销量的上升是由于某个季度某些商品类别的销量上涨导致的。

再来深究为什么第三季度和第四季度的销量主要贡献者类别5008168和50014815会在下半年出现大幅度增长。

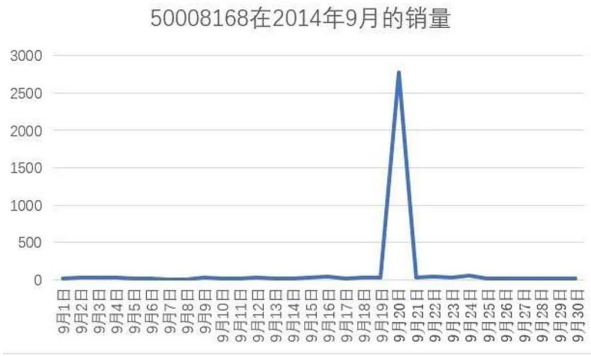
首先从类别5008168开始。搜集第三季度销量数据可以得到下图。



从图中可以看到，7-9月期间大部分时间销量都是比较平稳，唯独9月20日这天该产品的销量达到了2815。进一步搜集数据发现，是因为在2014年该产品的销量远远大于2012和2013年。



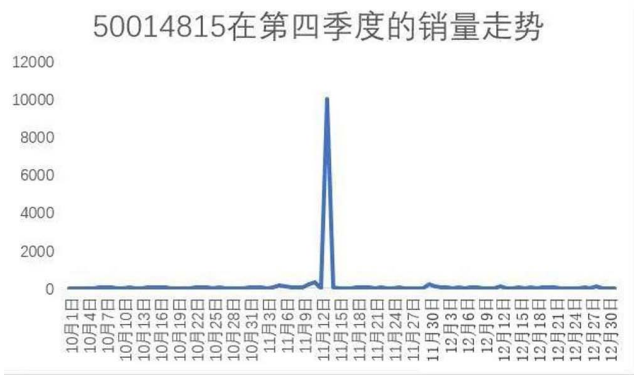
找出2014年9月该产品的销售数据，得到如下图。



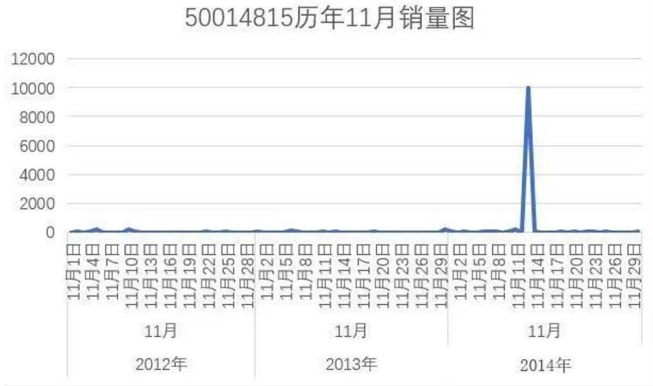
上图告诉我们，在2014年9月20号当天，508168的销量达到了2779。

由于当年当月的节日如中秋节在9月8号，教师节在9月10号，产生热销的原因没办法证实。但可以揣测是因为商家对该类产品做了促销活动导致的销量上涨，从而导致第三季度销量上升。

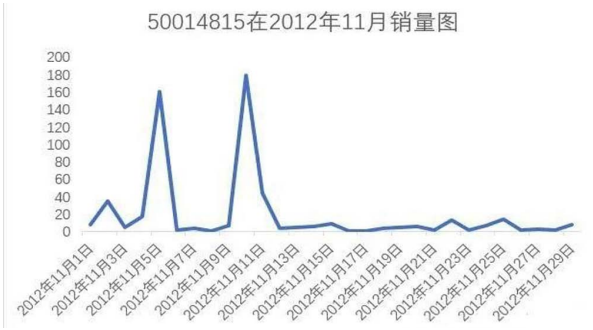
再来看类别50014815。

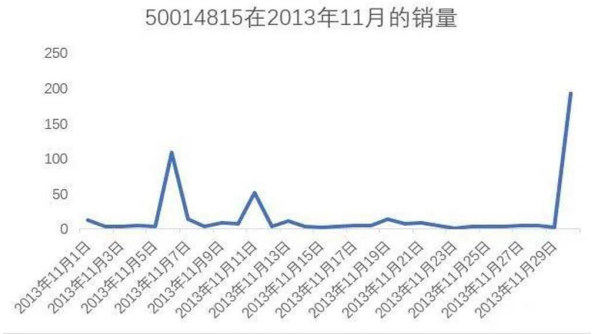


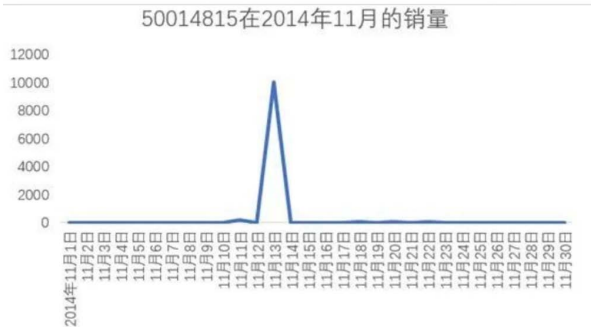
它在第四季度的11月份有一个显著的增长高峰。



数据告诉我们，它的增长主要来源于2014年11月13日的销量高峰，达到10029. 下面是该产品历年的11月销量图。





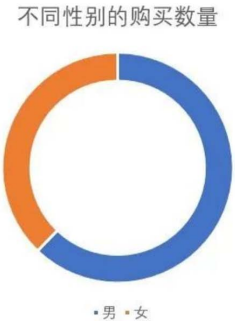


上面三个图我们可以看出，历年来11月的销售高峰并没有出现在双十一当天，而是2012年的11月10号，2013年的11月30号，2014年的11月13号。虽然2013年双十一那天出现了销量小高峰，但影响效果并不大。在其他日子出现销量大幅上涨，猜测是由于商家进行了其他促销活动，但缺乏数据支撑。

3. 性别对销量的影响

分析流程如下：

表1用If函数计算出成交单量，对表2用vlookup函数关联表1的购买日期，购买数量，商品大类，成交单量。清洗数据集并统计有效数据后发现用户里有406个女童，444个男童。所以男童用户比女童用户多。

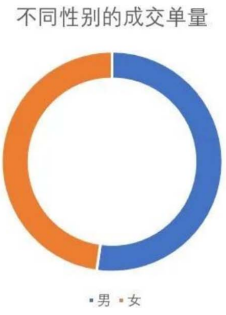


从上图得知，男女的购买比例为62%：38%。

显然男女用户的比例不足以造成如此悬殊的销量比例。

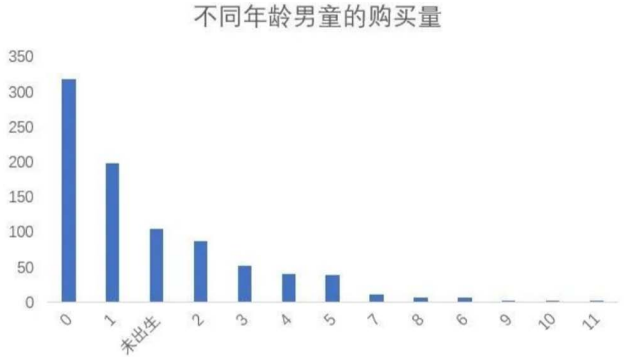
将购买量多维度拆解，可以得到：总购买量=成交单量*每单购买量

假设是因为男童的成交单量造成的。



男女用户比例跟成交单量比例是一样的，所以男童的购买量大于女童购买量可以说完全是因为男童的每单购买量大于女童的每单购买量导致的。

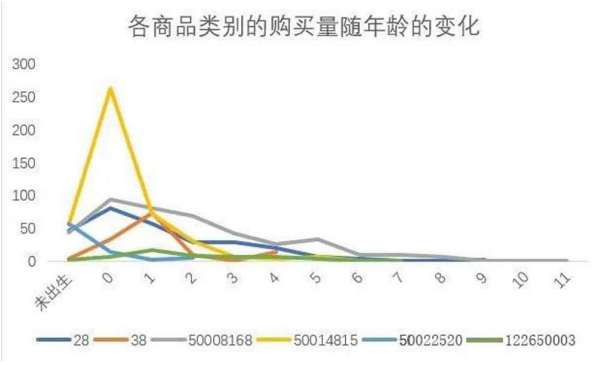
那又是什么年龄段的男童的每单购买量比较大呢？



可以看出，5岁以后的男童基本不再产生购买行为。主要买家为1岁以前的男童家长。

4. 年龄与销量的关系

购买量=各个商品类别的购买量之和



由上图可以看出，所有的类别的购买量随着年龄的增加都在下降。类别50022520从一开始就一直在下降，剩下的其他产品类别的趋势是先上升，幅度或大或小，然后再都下降。仔细看来，类别，50014815，50008168和28的销量高峰都产生在年龄为0岁，类别122650008和38的销量高峰产生在年龄为1岁的时候。

因此可以说1岁以后所有类别销量都在下降，可以猜测这些商品类别应该是适用低龄幼儿的产品。但不同年龄的销量高峰对应的产品类别不同，又说明这些产品的受众不同，应该采取分年龄营销策略。

结论

1. 依照现有数据2015年1，2月的销量相比往年没有下跌，反而比往年这两个月销量总和多。

2. 第三，四季度的销量相比于第一，二季度上升是因为个别商品类别购买量上涨导致，第三季度的增长主要是由类别5008168带来的，第四季度销量主要是由类别50014815带来的。而5008168的增长是由于在2014年9月20号当天，该类别的销量达到了2779；50014815的增长是因为2014年11月13日的销量高峰达到了10029，两者都发生在2014年。双十一购物节确实对刺激某些商品类别的销量有积极影响，但影响力度有限，有时候不及其他促销活动有效果。

3. 在销量上，男童大于女童，原因是男童用户的每单购买量大于女童用户的每单购买量。5岁以后的男童基本不再产生购买行为，主要买家为1岁以前的男童家长。

4. 对于1岁以上用户，所有产品的吸引力都在下降。但每个品类最大受众的年龄段不同，主要集中在0岁和1岁婴幼儿。

建议

· 优化影响单量的各个因素，如可以从产品，客户服务，退货款政策，广告等方面着手，在2014年的基础上进一步壮大用户基数，提高留存与复购率。

· 在下半年将不畅销产品类别与畅销品类捆绑销售，或者做加购活动，以带动整体销量。

· 调整产品范围，缩减适用于5岁后孩子的产品，集中供应这个年龄前的婴幼儿产品。尤其是要丰富1岁前孩子适用的婴幼儿产品，特别是男童，给顾客创造更多消费的机会。

· 升级改善1岁到5岁孩子适用的产品，可以从提升质量，捆绑营销，买赠等方面刺激销量。

· 采取分年龄营销策略，让每个品类精准辐射到对应的人群。

/ 基于大数据分析的电信用户赢回策略探究（以大学生市场为例） /

作者 / 中国联通武汉市分公司 CPDA数据分析师 韩露、郑茂、叶峰、刘宇、胡作梁、韩芹 编辑 / 协会会员处 李苗苗 日期 / 2020-05

摘要：通过参加中国数据分析专业委员会数据分析师（CPDA）培训并获得认证，使我们建立了良好的数据分析思维、掌握了一定的数据分析方法和技能，并结合实际工作进行相关的应用研究。本文通过对用户数据库中的数据进行挖掘分析，分析用户的离网现状、原因，挖掘出电信用户流失的影响因子，通过计算、分析，得出影响因子的阈值。并通过对大学生样本获取的被赢回顾客数据，探究电信行业大学生用户的赢回策略。

随着通信市场的饱和度以及电信产品与服务同质化程度不断提高，同时，运营商之间的价格竞争激烈，并且面临云联网技术的巨大挑战，离网用户的规模也日渐增多，移动产品作为电信行业的主要收入来源，在国家提速降费政策面前停滞不前，面临巨大的挑战。

因此，以数据分析视角研究电信行业离网用户赢回策略有助于明确离网用户的驱动因素及赢回的逻辑依据和动机因素，本文研究以大学生市场为背景，通过数据收集、处理与分析，从而指导电信企业更有效、更有针对性地制定大学生用户赢回策略，增加大学生用户的赢回效率。

1 概述

客户离网是全球各大电信运营商非常重视的话题，根据统计，美国电信行业的客户离网率达到了30%，欧洲则为25%。客户离网导致的直接后果就是公司利益的损失，因此对客户离网进行预警，能够对高潜在离网的用户进行挽留操作。如果存在高潜在离网用户，通过对该用户的历史行为进行分析，懂得用户真正的需求，给用户进行个性化的服务推荐，满足用户需求，可以减小用户离网的可能性。

本文研究数据处理主要可以分为下面三个部分：

一是基于用户行为、用户属性的离网预测，前期对电信大量数据进行预处理（样本采样、过滤，数据分类、归一化、离散化、特征降维等等）， 通过公式计算得出一系列规律，建立较为准确的模型和损失函数，使用正规化选择较好的算法模型，利用梯度下降算法对参数进行快速的确定，最后使用xgboost，将多个算法结合投票的得出结果。

二是基于得出的离网用户，配合各个不同营业员的特征参数，话术参数进行第二次回归分析将第一步得到的结果，作为第二步的参数，再次进行分类，利用 Softmax 回归迕进行分

类，得出结果。

第三是针对第一步和第二步的结果精准判断哪些是潜在用户，哪些是保有用户，哪些是离网用户，正确画出用户画像后，对比用户的购买行为，在第一层使用逻辑回归算法，第二层使用人工神经网络，从而实现精准营销。

2 数据处理

数据获取：在武汉联通，我们采用 HDFS 和 Spark 负责原始数据的存储和管理包括详细的通话记录单及宽带用户表。其中两张表均含有用户自身数据，包括年龄、性别、主套餐、融合 套餐、资费、基站、套餐使用情况、教育程度、通话时间、最大流量APP、流量使用前10名APP等。

数据预处理：（1）进行数据清理，对数据的唯一属性值进行删除（如身份证、姓名等）。（2）对数据进行缺失值填充，先对异常数据进行过滤，将其值变为控制，然后使用拉格朗日插值方 法对数据进行填充，使用回归分析法进行噪声平滑处理。（3）对特征值进行One-HotEncoding，使得我们能够处理非数值属性；在一定程度上扩充了特征；编码后的属性是稀疏的，存在大量的零元分量。（4）对数据进行标准化操作于每个属性，设minA和maxA分别为属性A的最小值和最大值，将A的一个原始值x通过min-max标准化映射成在区间[0，1]中的值x'，其公式为：新数据=（原数据-最小值）/（最大值-最小值）。（5）因为前期进行了One-HotEncoding，为了减轻维度灾难问题，对特征向量使用filter进行降维处理。Feature Engineering：特征工程模块将原始数据处理成和离网相关的结构化特征，用作分类器的输入，在这里，我们使用GMM和EM聚类方法。我们将用户的特征集进行划分，得到了三种行为分类：通信行为判别模型、交友圈与社交行为模型、业务质量感知评估模型。

Classifiers：利用分类器训练出来的模型预测未来有离网倾向的用户，按照离网倾向高低排名，根据这个名单进行个性化维挽。在分类过程中，我们整体的算法使用了投票的机制，运用多种机器学习算法，得出阈值，再对阈值迚行xgboost分类，从而得出最终的结果，在第一层，我们分别使用了SVM算法、随机森林算法、逻辑回归算法，最后使用决策树将三种算法的阈值迚进行分类，得到最后的分类结果。

最后将用户维挽的结果反馈到模型中形成闭环，不断提高模型预测容易维挽的离网用户精度。通过设置一个预警值来进行离网预警，使用随机森林画出用户特征值的影响程度，并找到最相关的特征变量。通过交叉验证，迚进行模型的优化，防止过拟合和欠拟合，模型融合可以比较好地缓解训练过程中产生的过拟合问题，从而对于结果的准确度提升有一定的帮助。用python中scikit-learn里面的Bagging来完成。

结合营销话术进行二次分类：本文最大的亮点就是在找到离网用户和即将离网的情况下，如何通过营销话术和用户行为偏好挽留用户，在这里，我们结合已经画好的用户画像，对我们的营业员数据和营销数据进行结合，再次进行新的一轮数据清理，使用人工审计网络的方法对每一项特征值计算得出相对应的权值，使用后向传播算法对其进行二次分类。

收集用户属性和偏好。要从客户的行为和偏好中发现规律，并基于此给予推荐，如何收集用户的偏好信息成为系统推荐效果最基础的决定因素。

找到相似的用户。当已经对用户行为进行分析得到用户喜好后，我们可以根据用用户喜好计算相似用户，然后基于相似用户进行推荐，这就是最典型的基于用户的协同过滤。最后采用皮尔逊相关系数或者余弦相似度计算用户的相似度。

计算推荐。基于用户对物品的偏好找到相邻邻居用户，然后将邻居用户喜欢的推荐给当前用户。计算上，就是将一个用户对所有物品的偏好作为一个向量来计算用户之间的相似度，找到K邻居后，根据邻居的相似度权重以及他们对物品的偏好，预测当前用户没有偏好的未涉及物品，计算得到一个排序的物品列表作为推荐。

初始化推荐列表，对列表迚行过滤、排名等处理，从而生成最终的推荐结果。

3 策略分析

（1）顾客赢回

顾客赢回是指企业采取积极有效的补救措施，重新恢复与流失顾客之间业务关系的过程，是企业重新激活与已流失顾客关系的过程。在电信行业，研究发现，流失顾客先前交易期关系的整体满意度显著正向影响赢回绩效，而关系时长尽管不显著但负向影响赢回绩效。

（2）赢回策略

赢回策略是指企业为赢回流失顾客所采用的营销手段和工具，其对赢回流失顾客至关重要。一般而言，赢回策略主要包括价格促销和关系投资，前者是指企业在某个特定时期通过降低某种商品的价格，或增加单价下商品商量的一种营销手段。后者是指企业通过投入大量的时间、精力甚至金钱等资源，设法越过关系门槛，取得顾客信任和情感依附，从而在企业与顾客之间建立起关系纽带，并达成长期合作的目的。

价格促销策略无法赢得顾客信任，但对顾客持续使用意愿仍具有一定效果。关系投资是顾客赢回的重要策略，且赢得信任至关重要。尽管在之前的营销经验中，价格促销都是有效顾客赢回的“利器”，但对于新时代的大学生而言，价格促销策略相较于关系投资策略效果已经变弱，且不能在流失客户中建立信任，因为电信运营商在赢回客户的管理过程中，应尽量减少使用价格促销策略，转变现有的营销理念，将赢回工作重心放在关系投资商，通过专业的顾客赢回服务团队，丰富的沟通工具，增加互动，促进价值共创行为或精心挑选赠品等，加强与流失客户的深层情感沟通，最终越过情感“门槛”而建立足够的信任，从而一贯的流失客户的青睐。另外，随着大学生消费支配能力的提升，其越来越关注使用体验，因为运营商应在套餐涉及、硬件及技术上加大投资，减少顾客因产品或信号不好而流失。

/ 通过消费者购药评论的关键词挖掘，浅析新冠肺炎疫情影响下，传统OTC医药零售市场的驱动力变化 /

作者 / 武汉CPDA数据分析师 朱科　编辑 / 协会会员处 李苗苗　日期 / 2020-05

背景

目前疫情对国内未来的影响，所涉及到的各行各业各层面渗透之广，摊开来分析，十分困难。因此，结合笔者所在行业，浅谈一点疫情影响所带来的传统OTC零售市场的驱动力变化，来了解一下未来医药零售行业可能的变化趋势。

首先，我们需要知道本次事件的背景影响有多深，从而判断其影响力的强弱。本次疫情已创造了多个世界性历史记录：

1、新冠肺炎为西班牙大流感以来，最大规模的全球性流行疾病。

2、新冠肺炎直接导致1930年经济大萧条之后，最严重的经济停摆，乃至衰退。

3、间接影响，自OPEC成立以来，首次石油期货价格跌至负数的奇观。

由于新冠肺炎的传播特征，及无特效治疗药物的干预。目前，隔离已成为最有效的预防手段之一。当社会缺乏广泛、高速地流动性之后，消费习惯的改变所带来的消费市场的驱动力的改变，已成为疫情之下的定局。关于这一点，也可结合2003年“非典”疫情所带来的对消费市场的变革来参考。

对于传统OTC零售行业来说，渠道流通+终端拉动是最为广泛、快速，有效的营销手段。但本次疫情下，人们居家隔离、减少外出，消费者从无法去线下实体药店购药，转变为线上购药的习惯，导致医药线下终端市场受到巨大冲击，而由于医药行业在线上的经济规模远低于其线下规模，线上增长额低于线下下降额，从而直接导致OTC零售医药市场销售金额的下降。

线上购药迅速成为疫情期间主要的购药渠道之一，而线上购药方式受消费者自我判断的影响较大。当消费者习惯了线上购药的消费方式之后，传统的OTC零售市场的驱动力影响会逐渐削弱。因此，分析消费者对家中常备医药产品的喜好因素，成为分析产品驱动力的其中一条途径。

目的

通过分析消费者对不同产品的评论，拆分高频词汇。找出消费者在意、关注的重点，提供产品分析相关的信息。从而探索医药零售行业未来的发展方向。

分析路径

通过python等数据工具获取自1月下旬疫情爆发以来网络中，消费者对各治疗领域的药品消费评论，进行有效关键词的拆解。我们可以获得以下词云结果。

我们选取本次疫情最受影响的感冒及止咳类产品，进行词云分析如下：

感冒类OTC零售药品

止咳类OTC零售药品

小结

在感冒类OTC药品市场中

① 消费者的关注点主要集中在传统的物流、效果、价格上。

② 品牌是最大的拉动力之一（高频词汇如信赖、正品、牌子）。

③ 受疫情的影响，抗病毒出现高频（高频词汇如疫情、抗病毒）。

④ 消费者已经开始注重家庭常备了（备用、常备、预防）。

⑤ 专业专家及医生的背书作用较疫情之前更加明显（连花清瘟与香雪抗病毒销量增长较快）。

在止咳类OTC药品市场中

①与感冒药不同的是，高频关键词中出现咳嗽、枇杷膏，说明消费者对“止咳=枇杷膏”有较深的消费者认知。

②念慈庵的评论中，止咳的出现频率较高。说明消费者对该类品牌的效果认知较为强烈。

③康隆强力枇杷露的评论中，多次出现医院词汇。可能存在医院推荐的产品目录之中。

结论

结合以上的细分小结，我们不难发现，在终端零售药店的店员推荐失去以往的推荐途径和力度之后，消费者的自我认知与选择，成为了疫情之后，OTC药品销量的主要影响因素。

那么，对于医药生产企业来说，如何从传统的线下渠道通货———零售终端药店店员推荐的模式，转变为线下流通与线上品牌力影响双管齐下，会成为未来医药行业增长的常态营销策略。

可预测的是，在医药行业，品牌力影响与专家医院背书，会成为消费者选择的侧重方向。有品牌影响力、有口碑的知名药企的发展机遇会增大，而高毛利不知名的小型医药公司的生存空间会逐步缩小，生存压力增加，但也并非全然坏事，下面我会提到。

同时，疫情影响会加速线上化的交易频次与交易量，而医药行业在线上平台的发展相对于其他零售行业较晚，也可以说，是相对机会空间更大。

再次，医药线上化的增长，会促进多元化的发展。传统线下医药零售的分类，在融入线上平台之后，由于线上用户人格的细分化程度远超于线下单一化的零售药店店员推荐，线上用户人格会越来越真实地反应用户的真实需求，对产品的细分化会更加突出与明显。同时，医药行业不同与其他快消品行业，医药行业在兼顾消费者自身需求喜好的同时，还有着严格的医疗科学划分细则，那么对于不同体量的医药企业来说，其原有的线下渠道营销模式则需要发生改变，以此来适应新的市场格局的到来。

对于大型医药制药企业来说，具有巨大体量既是企业实力的体现，同时也伴随着巨大的市场挑战。如何丰富完善自身的产品结构，打造一整套的产品“帝国”，从而360度满足消费者一站式购药的需求，同时针对不同消费人格的用户，以及不同的治疗领域、症状，做到面面俱到，是大型医药制药企业需要思考的方向。在拥有企业核心竞争力的基础上，以品牌化效益逐步进阶到整个品类，成为消费者第一联想对象。

而对于中小型医药制药企业来说，规模及资金的限制，制约了该类企业走广而全的路线。而疫情在压缩传统高毛利的产品生存空间之外，同时带来了另一种机遇。多元化的加速，也会伴随着幂率化的增长。原本需要与众多大型医药制药企业的竞品在同一条“独木桥”上竞争的情况，可能发生改变。单一化的市场会逐步分化为无数个小的个性化市场。中小型医药制药企业的机会，则是抓住其中的部分市场进行集中攻坚，不失为一条新的道路。

以上，是笔者对这次新冠肺炎疫情影响下，传统OTC医药零售市场的驱动力变化的粗浅分析。

/ 基于数据分析法的运营商项目采购中供应商选择研究 /

作者 /武汉 CPDA数据分析师 帅旗 张珏 黄卓 吴明 编辑 / 协会会员处 李苗苗 日期 / 2020-05

一、研究背景

近些年运营商竞争趋于白热化,电信行业已经告别了高利润的时代，运营利润的下降，驱使运营商一方面重视新业务的开发和新技术的应用，以便提供更具竞争力的通信服务产品。另一方面更加关注成本，成本的控制将对利润率水平产生重大影响。因此，为了降低采购成本，发挥规模采购的优势，运营商越来越多的项目采购模式由过去的分散采购转向了集中采购,这就对项目的采购管理提出了更高的要求。供应商选择不仅关系到项目成本，也将影响到运营商竞争力的提升和运营目标的实现，如何进行科学供应商选择已成为运营商需要优先考虑的关键问题。

二、运营商项目采购中供应商选择原则

供应商的选择是采购过程中的重要环节,如果供应商选择不当，不仅影响后续网络的发展,同时还可能影响现有网络质量，而降低了运营商的整体竞争力。在实际的供应商选择过程中，该根据不同的产品特点及具体的项目情况选择不同的原则和标准。

一是核心竞争力，选择具有较强核心竞争力的供应商，且与供应商之间形成一种长期合作的比较稳定的战略伙伴关系,共同推进整个产业链的发展。

二是正确的价值观，择强供应商，绝不以牺牲产品质量为代价来降低产品成本，而更加关注通过提高技术水平和营销策略来达到降低采购成本方法。

三是较强的技术实力和自主研发能力，选择该领域内的技术引领者或该领域内比较公认的技术实力和自主研发能力较强的供应商。

四是市场的快速响应能力，供应商具有快速的新 产品、新功能的研制和开发能力。

五是售后服务能力，选择具有较强售后服务能力的供应商，具有完备的服务体系，提供全国范围内的快速响应服务。

六是设备应用广泛，选择该领域内设备应用广泛的供应商，该类供应商的设备往往更加稳定，设备兼容性更强,更能适应复杂的现网环境，具有更加丰富的网络建设经验。

三、建立供应商选择模型

（一）确定供应商选择方法

大数据时代，运营商因其特殊的行业特征，在进行供应商选择，既要考虑货源的稳定性和及时性，以确保生产的连续性，又要考虑其他众多的影响因素，如供应商的产品质量、生产能力、产品交货期、产品结构、营销情况、产品价格、公司信誉以及企业的市场影响力等。在具体的供应商评价指标筛选中，应该从使供应链绩效最大化的目标出发，按照指标的设计原则，来确定供应商的评价指标。本文从运营商项目采购中供应商的影响因素出发，建立供应商选择决策的评价指标体系，采用既有定性分析又有定量分析的层次分析法（AHP）确定出各指标的权重，然后再利用综合评价法建立综合评价模型。

（二）设立供应商选择的综合评价指标体系

评价指标体系是指由多个不同角度、不同层次的具有内在联系和特性的指标所组成的有机整体。本文根据工作实际，结合实地调研和深入分析得出影响供应商选择因素主要可以分为产品架构、系统集成性、可持续发展性、行业经验、研发能力、价格因素等六大类。

（三）运用层次分析法（AHP）确定评价指标权重系数

层次分析法（AHP）是20世界70年代，由美国运筹学家T.L.Saaty提出的一种多准则决策方法，常被用于多目标、多准则、多要素、多层次的非结构化决策问题，其主要分析步骤包括：建立层次结构模型、构造判断矩阵、层次排序及一致性检验。

1、建立供应商选择的层次结构模型

建立层级结构模型即根据决策目标将影响因素指标体系按照因果关系分成多个层次，如目标层、准则层、方案层等。本文结合供应商选择综合评价指标体系，将供应商选择的层次结构模型分为目标层A、中间因素层Bi和方案层Ci，层次结构模型如图1所示。

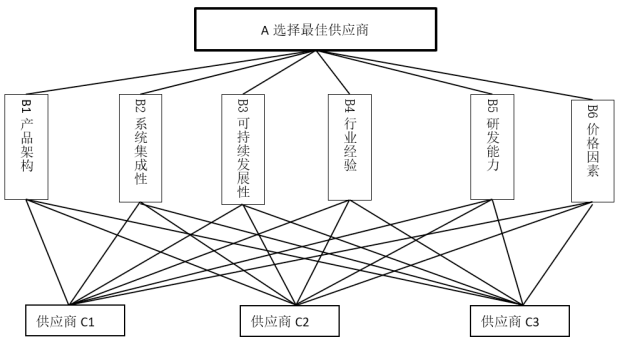


图1 运营商项目采购中供应商选择的层次结构模型

2、构造判断矩阵

根据T.L.Saaty提出的1-9标度方法，按照层次结构模型逐层构建判断矩阵。本文基于专家评定和市场调研确定出各要素权重，并获得7个判断矩阵，如表1-7所示。

表 1　判断矩阵 A-Bi

A	B1	B2	B3	B4	B5	B6
B1	1	2	1	3	1	1/2
B2	1/2	1	2	4	1	1
B3	1	1/2	1	5	3	1/2
B4	1/3	1/4	1/5	1	1/3	1/3
B5	1	1	1/3	3	1	1
B6	2	1	2	3	1	1

表 2　判断矩阵 B1-Ci

B1	C1	C2	C3
C1	1	1/3	1/2
C2	3	1	3/2
C3	2	2/3	1

表 3　判断矩阵 B2-Ci

B2	C1	C2	C3
C1	1	1/4	1/3
C2	4	1	4/3
C3	3	3/4	1

表 4　判断矩阵 B3-Ci

B3	C1	C2	C3
C1	1	3	2
C2	1/3	1	2/3
C3	1/2	3/2	1

表 5　判断矩阵 B4-Ci

B4	C1	C2	C3
C1	1	1/3	4
C2	3	1	7
C3	1/4	1/7	1

表 6 判断矩阵 B5-Ci

B5	C1	C2	C3
C1	1	1	6
C2	1	1	6
C3	1/6	1/6	1

表 7　判断矩阵 B6-Ci

B6	C1	C2	C3
C1	1	1/5	1/7
C2	5	1	1
C3	7	1	1

3、层次排序及一致性检验

表 8 随机一致性指标 RI 的数值表

阶数	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
R.I	0	0	0.52	0.89	1.12	1.26	1.36	1.41	1.46	1.49	1.52	1.54	1.56	1.58	1.59

化后）即为权向量；如果不能通过，需重新构造判断矩阵。本文采用根法求解，各判断矩阵的特征向量W和λmax计算结果如表9所示。有四个矩阵排序一致性指标CI为0，具有完全的一致性，另三个矩阵排序一致性指标CI接近于0，具有满意的一致性。所有判断矩阵的一致性比率CR均小于0.1，满足一致性要求。

表 9 各判断矩阵的特征向量 w 和 λ max

A	w	B1	w	B2	w	B3	w	B4	w	B5	w	B6	w
B1	0.183	C1	0.167	C1	0.125	C1	0.545	C1	0.263	C1	0.462	C1	0.078
B2	0.192	C2	0.500	C2	0.500	C2	0.182	C2	0.659	C2	0.462	C2	0.435
B3	0.190	C3	0.333	C3	0.375	C3	0.273	C3	0.079	C3	0.077	C3	0.487
B4	0.053												
B5	0.152												
B6	0.230												
λ max	6.489		3.000		3.000		3.000		3.032		3.000		3.013
CI	0.098		0.000		0.000		0.000		0.016		0.000		0.006
CR	0.077		0.000		0.000		0.000		0.031		0.000		0.012

底层指标Ci对目标层A的总排序如表10所示。在有多个供应商备选时，可依据表10的总排序权重作出选择。

表 10　C 层对 A 层的总排序

	B1	B2	B3	B4	B5	B6	C 层总排序
	0.183	0.192	0.042	0.304	0.122	0.171	
C1	0.167	0.125	0.125	0.125	0.125	0.125	0.260
C2	0.500	0.500	0.500	0.500	0.500	0.500	0.427
C3	0.333	0.375	0.375	0.375	0.375	0.375	0.313

（四）综合评价分析

由此可以得出以下结论:应首先选择供应商C2，其次是C3，最后才考虑C1。在供应商选定以后，应进行跟踪动态考评，对自觉提高产品和服务质量的供应商，企业应该给予鼓励，以提高供应商的积极性。对那些偷工减料、不负责任的供应商应予以相应的惩罚。并随着企业的发展变化对判断矩阵的权重予以调整，以适应企业各个时期对供应商的要求。

四、结语

基于数据分析方法，本文利用综合评价和层次分析建立了基于数据分析法的运营商项目采购中供应商选择模型，为运营商提供了可供参考的量化评价方法，在一定程度上可以拓展相关领域的分析视角。

/ 运用数据分析方法解决潜在医药品品种和用药人群分析的研究 /

作者 /武汉 CPDA数据分析师 杨红鹰　编辑 / 协会会员处 李苗苗　日期 / 2020-05

身为公司市场研究员，今年参加了中数委CPDA数据分析师培训认证，通过培训全面、系统地学习了数据分析的方法、技巧、工具等知识，不仅顺利获得了CPDA数据分析师证书，还掌握了实际应用的技能，建立了行之有效的数据分析思维。并结合实际工作进行相关的应用研究。本文通过运用数据分析方法解决潜在医药品品种和用药人群分析，实施精准营销，扩大市场份额。

随着“三医联动”系列政策的落地，医药行业迎来了前所未有的巨变，格局正在重塑，市场竞争变得空前激烈，医药企业无一例外面临严峻考验。在此情势下，企业如何摈弃经验主义，深度了解市场，如何从科学的角度寻求新利润增长点、如何将有限的资源产出最大价值等问题，都急需解决。

基于大数据分析可使用不同分析方法和模型，为产品研发分析、销售预测、市场洞察分析、产品和用户分析、生命周期管理提供有效参考。我通过将数据分析思维和方法融入实际市场研究中，从而获得了较为理想的效果。在此，我想先分享用MECE来提升品种筛选参考价值、借用户行为分析助力精准营销的方法。

一、用数据分析方法解决品种筛选难题

市场是主体，客户关系是关键，而产品是目的，在产品策略研究期，筛选更具潜力的产品是极为重要的一环。我们更需要发现具成长空间的药品。

因此该问题可转化为：寻找哪些药品具有较大成长空间和市场刚需。因品种评估向来标准不一，效果也会云泥之别，所以用科学建模的方法进行，不失为一种具有可提升分析结果参考的价值，并减少商业策略风险的有效方式。

根据MECE方法论，我需穷尽所有相关且独立因素，并找到主要矛盾，再将其转化为可用数据分析实现的问题。据此，我将主要与之相关的核心因素尽数列出，并选取代表性因素，对目标建立多维评价体系（建模），赋权后再综合分析，最终将分析结果结合政策走向等因素，来确定结果。

考虑到药品评估可从“评价该产品是否健康发展”的角度来挖掘，因此增长百分比为一个评判指标；但若市场份额足够大，增长也不会显著，也可能是这品类市场趋于饱和，或这一品类已处于成熟期，因此再引入销售增长率来评价，即可预判药品在品类格局里的短期趋势（动态指标）。

除此之外，该类产品还应具有一定市场份额，确保其市场刚需和前景，因此再纳入市场份额，而集中度事关竞争程度，能提示介入风险的大小，也将其纳入，这样主体模型基本可确定了。

综合以上指标，对各终端化药和中成药A品类产品展开挖掘，这里需提取至少3年数据；

大类	亚类	市场份额	市场份额增长	增长率	CR8

按化药和中成药类别，我分别对相关品类的市场份额、市场份额同比增长、销售增长率、市场集中度展开建模分析，取调数据（因已有购买权威数据库，所以对数据不再出处理），并对影响产品健康成长的关键两项加权，得出下表概况（数值均已做修改或删除处理）：

2016-2018年**终端用药综合分析

大类	亚类	市场份额	市场份额增长	增长率	CR8
化药		1	3年持续增长	3年持续增长	%
		2	2年持续增长	2年持续增长	%
		3	3年持续递减	近2年持续增长	%
		4	震荡回稳	3年正数增长，震荡提升	%
		5	3年显著提升	3年显著提升	%
		11	近3年正数增长	近2年平稳 ，略有回落	%
		10	第3年略有回升	第3年稍有回升	%
			3年递减	3年正数增长、震荡	%
		8	增长稳定	3年增长稳定	%
中成药		1	3年正数增长	3年正数增长	%
		2	第3年显著回升	3年正数、持续、显著提升	%
		3	增长持续提升	3年正数、稳定、显著提升	%
		6	3年持续提升	3年正数、持续、显著提升	%
		8	3年持续回升、提升	3年持续回升、提升	%
		9	3年持续、显著回升	3年持续、显著回升	%

因在医药行业，药品受政策影响极大，所以基于此结果，再用医保用药、基药目录、合理用药目录、渠道等对结果进行了第二轮筛选，使结果更具科学参考价值。

这种方法得出的分析结论，源于关键核心的因素，能有效的反馈该产品在市场的竞争力，成稿获得了品牌经理的较高认可。

二、数据分析实现精准营销——用药人群画像

一个产品若想得到广泛应用，受众分析必不可少。产品经理需要懂客户，除了购买行为外，还需了解客户深层的动机与心理，为其实现精准营销提供细分领域的深入洞察。

客户画像是受众分析中的最有效方式，能让客户经理对目标客户了然于心，在此我尝试通过数据特征的变化来实现。

基于现有数据品类，我考虑需要了解产品的用户人群、变化和用户的行为特征，这其实分别可与市场份额、市场份额同比、及销售增长率相对应，再结合药品的具体属性、功效，基本可实现对细分市场的用药人群画像。

在此依然引用数据库的二手数据来做剖析，我主要提取了市场份额、计算市场份额增长率、销售增长率三项指标进行统计。

**终端A品类化学用药市场份额同比增长

亚类	2016年变动	2017年变动	2018年变动
T1	-1.75%	-0.88%	0.92%
3	3.31%	-2.80%	-0.24%
1	-0.88%	-1.85%	-3.46%
6	-2.17%	6.64%	1.34%
T2	-4.71%	6.69%	7.36%
5	2.46%	-3.09%	-3.90%
11	1.54%	6.82%	-4.96%
2	-0.42%	2.51%	0.00%
10	8.97%	12.66%	-2.25%
4	-7.14%	-10.99%	-7.41%
0	6.25%	5.88%	3.70%
7	6.45%	15.15%	-2.63%
8	22.22%	45.45%	-43.75%
9	0.00%	0.00%	0.00%

**终端A品类0类化学药市场份额同比增长

0	6.25%	5.88%	3.70%

***终端A品类0类中成药市场份额同比增长

亚类	2016年变动	2017年变动	2018年变动
0	-3.90%	-3.90%	6.17%

A品类用药**端销售增长率

亚类	2016年	2017年	2018年
0	0.04%	3.97%	12.29%

从中可以看到，0类化药市场同比逐渐下降，而中成药市场平稳且提升，对应该品类的销售增长率则是显著提升，我们可以猜测，消费者最初偏重于选择治疗属性较强的化学药，可能是因为化药见效快，但近来用药习惯已在改变；

从表一还可以得出，哪几类疾病属于慢病，后期市场相对稳定，可关注慢病管理终端；哪几类治疗属性偏弱，但市场表现持续提升，反映了大众保健养生意识，一方面应和此类高毛利相关；（略）

因药品名称限制（对应疾病和病症反映），在此不能有效呈现。但在实际操作中，我们已通过类似此分析方法，得出针对性更强，更完善的画像结果。

此分析报告的完成，打破了我们对细分市场单方猜测，且难以验证的困境，为该产品实施精准营销提供了有效参考依据。

在大数据时代，人工智能、区块链飞速发展，我们甚至迎来了5G时代，我将利用现有知识储备、多年实战经验和数据分析思维的养成，探索将大数据分析应用于更多实际案例中，尝试去解决企业经营的难点、痛点，最终实现提升企业经营效益的目的。

/ 钟南山院士团队新论文方法的启发 /

作者 /广州 CPDA数据分析师 刘程浩　编辑 / 协会会员处 李苗苗　日期 / 2020-05

在5月12日，钟南山院士的团队在医学杂志JAMA-Medical发表了最新研究成果“Development and Validation of a Clinical Risk Score to Predict theOccurrence of Critical Illness in Hospitalized Patients With COVID-19”。我翻译成中文是“COVID-19的住院病患产生重症的临床风险评分预测方法的开发和验证”。

看到新闻后的第一时间我就登陆了JAMA对论文进行阅读，费了一些劲儿才把文章看完。说实话这辈子如果不是因为拜读钟院士团队的文章，恐怕到老都不会接触那些医学专业术语，像是neutrophilia（中性白细胞增多）、coagulopathy（凝血病）……通篇阅读完之后我有以下几点感触和启发，这些想法也增进了我对数据分析认识上的深度解读。下面我就来简单说下我的这几点感受。

同一场疫情下，不同模型所产生的认知差异

今年年初COVID-19疫情汹涌而来，而且又还是一个新的未知病毒作祟，人们都很紧张。普通老百姓最担心的是疫情的病毒的传染能力、如何避免传染、是否高致死等关系切身生命安全的问题。

大家回忆一下疫情初期朋友圈里热衷于探讨“繁殖率R0”的文章里面，民间有一些机构和学者研究计算的值也各不相同。有人说是2.4，也有人说是4.5……。然而很多人只记住了“一个人传几个人的繁殖速度”，却没有看到这个指标背后的很多基础假设和计算背景。不过这也不怪大家，因为我们都是非医学专业的。

这些各种版本的R0在民众惊恐心理支配下，被很多不良的自媒体拿来各种歪曲炒作。我在认真学习了经典的传染病模型知识之后，看了很多搞文艺，或者号称研究区块链的自媒体也在发表文章，通过对R0来预测疫情了！看着半桶水哐当哐当地论调在博眼球，还没读完我就很想笑，到后来就干脆不想看了。老百姓对模型的研究理解，大致就是到这个层面。

钟南山院士团队第一篇关于预测感染人数的论文，是应用了改良后的经典传染病模型SEIR。刚好我在这次疫情期间也对这个模型进行了研究，通过对模型的研习，我发现定量的分析能够以一个更高的视角看问题。

比方说深度理解疫情下那么多防控措施背后的为什么，进一步看懂社会医疗资源为什么进行这样或那样的分配……等等。这里有一个基本的逻辑关系，就是确诊人数的增长速度和模型的预测进行比对，可以用于校验前期模型的合理性。如果确诊人数增长过快，那么说明原先模型的假设以及参数已经失效，相对应的原有防控措施需要更加严格。而如果确诊人数增长符合模型预测，或者趋缓，甚至下降，那么就说明原有的防控措施是有效的。

以经典的传染病模型SEIR为例，里面有2个重要参数：感染者有效接触易感染人群的速率λ、感染率β。而大家在整个疫情期间所经历过的各种封闭管制，其实都是在控制不让这2个参数继续变大。

例如大家熟知的停止举办大型群聚活动、春节假期延长、公共交通停运、小区封闭式管理、戴口罩出门……只要不主动的聚集，控制好人员的流动，那么感染速率λ和感染率β就能控制住或者降低。当然随着疫情的发展，管制措施的逐步调整，对这两个参数的估算也是不断地动态变化的。

另外单独说一下模型中的这个参数：感染者有效接触易感染人群的速率λ。这次疫情中，包括国外抄作业都抄错的意大利等西欧国家，都发现单纯的封闭禁足是难以控制疫情的发展。因为家族传播在疫情扩散中扮演着非常重要的角色。要知道一家人中只要有一个被传染了，那么在居家禁足令的实施下，整个封闭的家庭是很容易全部“中招”的。这个时候λ在一个家族传播路径中就会变得很大，按照中国人的居家特点，λ=5或者=8都是正常不过。这样一来，也就难怪感染人数的发展非常快。另外这次疫情发现家族传播的病征都是普遍轻度或中度为主，如果这个时候能够识别和进行隔离，那么将会有力的阻断病毒的传播和感染者病情的发展。

面对家族传播的这一特点，知道了一个家庭某个人确诊，就可以预算出全家人所需要的病床资源，举一反三进行测算，更可以估计出某个区域未来潜在的病床需求数量。可问题是医院的病床远远不够，怎么办？这个时候一些大型的公共设施，例如体育馆、学生宿舍……被划拨出来建设成“方舱医院”。由于方舱医院改建难度低，模块化的组合也可以根据家族传播路径成片区的安置病人。

事实证明，方舱医院的投入使用极大的缓解了病床资源的紧张局面，及时的治疗有效的降低了轻症病患向重症患者的发展，同时也有效隔离了感染者更多的接触易感染人群。这样一来，家族传播在疫情中解释了λ的结构变化；而方舱医院的投入，单从控制家族与家族间传染的视角来看，反过来有力的控制住了λ。

以上是从一个较为宏观和定量的视角来感知模型（通过预测感染人数来整合社会医疗资源）。

钟南山院士团队这次在JAMA-Medical杂志上发表的论

文，则是从一个微观和定量视角来应用模型。

简单的对这个模型的应用做下介绍：

模型和工具开发的目的，是为了尽早地识别和预测COVID-19感染者演变成重症的可能性，因为它有助于提早准备治疗资源和安排适当的治疗措施。

为了能更好的说明这个问题，我引用了一下百度百科对ICU病房的介绍。因为ICU是重症新冠患者治疗的重要医疗资源。重症患者一般都是需要ICU即重症加强护理病房（IntensiveCare Unit）来开展治疗的，ICU因此也叫作加强监护病房综合治疗室。在ICU里面治疗、护理、康复均可同步进行，为重症或昏迷患者提供隔离场所和设备，提供最佳护理、综合治疗、医养结合，术后早期康复、关节护理运动治疗等服务。

本次的疫情在武汉的一些重症患者还要进入到“火神山”、“雷神山”治疗。在ICU里面，每床位的占地面积为15-18㎡，床位间用玻璃或布帘相隔。ICU的设备必须配有床边监护仪、中心监护仪、多功能呼吸治疗机、麻醉机、心电图机、除颤仪、起搏器、输液泵、微量注射器、处于备用状态的吸氧装置、气管插管及气管切开等所需急救医疗设备。在条件较好的医院，还配有血气分析仪、微型电子计算机、脑电图机、B超机、床旁X线机、血液透析机、动脉内气囊反搏器、血尿常规分析仪、血液生化分析仪等。

以上的医疗设备资源的高度利用，充分说明了普通病房和ICU病房的在硬件上的区别。而一家医院或者一个地区的有ICU病房的医院数量是有限的，如果大家都害怕自己万一感染确诊后会发展成重症，都往有条件的医院挤，那样不仅容易造成医疗资源的挤兑，也会造成医疗资源的不合理分配。可如果说一开始确诊时没有识别出短期内会发展成重症的“高风险”病患，把他放在普通病房接受一般的治疗，则容易错失治疗的最佳时机，原本有机会康复的病患却因此造成了悲剧。所以，一旦新冠病患能够在确诊的同时，能预测出短期内是否会发展成重症就尤为重要了。

可问题摆在这里，一个新冠患者的检测指标多达72项，哪些指标能对这个“潜在风险”起到关键指示作用？ 如果能把它们提炼出来是非常有价值的。因为并不是所有的医院都能检查这么全面，也并不是所有的医院的医生都参加了这次疫情的“湖北会战”，有和新冠病毒的实战经验。拿到检测结果却难以做结论的医生肯定不在少数。另外，这些关键指标的数值要严重到什么程度，才容易判定或者识别成高风险患者，这就需要有一套模型来测算。如果模型搞的很复杂，要求医生除了有专业知识和丰富的临床经验外，又还要懂大数据的数学模型算法，那实在又太强人所难。

因此，模型工具的简单化和亲民化是第二个现实的问题。所以，这次钟南山院士团队所开发的工具，解决的就是以上两个比较微观的定量难题。

我看了一下钟院士团队的最终成果，挺方便实用的。就是在下面的地址中，http://118.126.104.170/ 输入10个关键的检测指标，然后点击测算就可以了。我尝试填写了一下，结果显示我属于“低危人群”

模型选择上的原则——可解释优先

曾经大数据很火的2016年，业界对各种神经网络的算法尤其热捧。大家写论文也好，出方案也好，如果没有用上神经网络或者深度学习，好像就是落伍了一样。当然，我也没有免俗，人非圣贤嘛！毕竟神经网络算法最大的优点摆在那里，就是在分类算法中识别率很高！但是神经网络或者深度学习有个很大的问题一直无法彻底解决，那就是很难解释参数和超参数。也就是说这些参数或者超参数的现实意义是什么（物理解释）讲不明白。比方说网络结构中为什么隐藏层是5不是4，如果是5能代表什么现实意义？学习率为啥要设定为0.01不是0.015?总之很多地方谁都不能说得明白，所以这类算法也成为“黑盒子”。

与“黑盒子”难以解释相对应的是可解释的算法——统计模型。统计模型中最经典的则以各种线性模型以及可做线性变换的曲线模型为代表。不过线性模型有个最大的问题，就是拟合偏差比较大。而且我还是在学生时代，就发现很多经典的线性模型偏差太大！说到底因为现实的场景往往不是线性的结构。还有很多大数据相关的教材、讲课视频，入门的算法都是线性回归；当介绍完线性回归之后，立马就跳入了比较复杂的算法中去了。例如神经网络中计算梯度下降的公式中，会涉及到线性方程的计算部分。因此讲线性回归是为了引出“梯度下降”。不过这些跳跃经常让人难以适应。但线性模型最大的优点就是可解释性非常强。

举个例子来说

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

这个模型再简单不过了，模型的解释也可以非常的“白菜

话”：

· 如果|β1|>|β2|，说明自变量x1比x2重要。因为x1的变化引起因变量y的变化，要比x2更大。反之亦然；

· β0：就是说所有自变量啥都不干（不取任何值），因变量y本身就会体现出的平均水平。例如研究价格和销量关系中的“刚需购买”；

· β1：假设按住自变量x2不动，自变量x1平均每变化β1个单位，因变量y就会平均变化1个单位；

· β2：假设按住自变量x1不动，自变量x2平均每变化β2个单位，因变量y就会平均变化1个单位；

· 如果β1非常接近于0，那说明这个自变量x1对整个模型而言没有啥用处，因为无论它取什么值，和一个无穷小的数β1相乘，结果还是无穷小。

· ……其他

以上，尽管以线性回归为代表的统计模型可解释性强，但一直以来很多人都没将之纳入法眼。一方面是太简单了，体现不了“学术水平”；另一方面就是和“黑盒子”相比偏差大了些，怕同行取笑或者被导师打回去文章重写。我也自我批评一下：我曾经参加论文答辩时，看到有同学3万多字的论文中，居然只用了一个多元线性回归时，还一度觉得他不会是太偷懒了，拿个这么简单的模型来凑字数吧。不过这次看了钟院士团队的论文后，我倒是非常的受启发。

一、虽然线性模型本身拟合偏差比“黑盒子”大，但是它有很多高级版本的应用场景，你不知道而已。

就拿特征工程中，高价值的流程节点是如何解决共线性的问题。如果解决了共线性，在保证不降低预测精度的同时，一方面可以降低过拟合的风险，另一方面提取的特征少了（自变量），模型的计算量和模型对计算资源的耗费，将会大幅降低。本次钟院士团队采用的是广义线性回归中的LASSO回归（Leastabsolute shrinkage and selection operator），从72个检测指标中提取出了19个和COVID-19重症高度相关的指标；接下来在Logistics回归模型中，再次提取出了10个关键指标，作为重症患者的独立统计预测因子，最后纳入到风险评分工具中。

具体的指标见下表（医学术语翻译的可能不够准确，请学医的读者见谅）

LASSO 回归识别的 19 个指标	Logistics 回归提取的 10 个指标
CXR 异常、年龄、是否存在武汉暴露、初次和最高体温、呼吸频率、收缩压、咯血、呼吸困难、皮疹、意识不清、并发症数、慢性阻塞性肺疾病（COPD）、癌症、氧饱和度水平、中性粒细胞、中性粒细胞淋巴细胞比例、乳酸脱氢酶、直接胆红素、肌酐水平	CXR 异常、年龄、咯血、呼吸困难、意识不清，并发症数、慢性阻塞性肺疾病（COPD）、癌症、中性粒细胞淋巴细胞比例、乳酸脱氢酶、直接胆红素

在这里说到的LASSO回归，就是线性回归模型中，若存在多重共线性场景下，选择显著影响的自变量的方法。在这里论文并没有直接拿它来做COVID-19检测指标的回归分析，而是通过LASSO来做了一次特征工程，过滤掉了很多共线性和不显著影响的自变量指标。 这点确实让我眼前一亮。说实话，很多经典的非监督学习，例如相关系数矩阵、主成分分析法PCA、聚类分析、K-MEANS，逐步回归……都可以实现识别多重共线性的问题。但问题是你识别出了一组高度相关的指标，之后呢？比方说你发现了X1，X2，X3存在共线性问题，你该挑选哪个？过滤掉哪些个？这就是选择上的难题！LASSO回归刚好解决了这个难题。有兴趣的读者可以继续阅读LASSO回归的原理介绍和在钟院士团队论文中的应用。

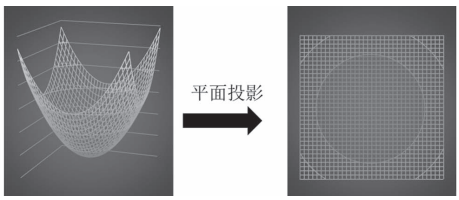
LASSO回归的原理，其实不是很复杂，之所以说它是线性回归的高级版本，实际上是有原因的。LASSO回归是Elasticnet模型家族理论中的一个特殊场景，它有一个同胞兄弟，叫作岭回归（Ridge回归）

一般的线性模型，为估计出模型系数的β向量，是对以下的残差平方和RSS求极值而来。

$$\min_{\beta_0,\beta}\left[\sum_{i=1}^{N}\left(y_i-\beta_0-x_i^T\beta\right)^2\right]$$

其中，N=样本个数

因为上面的残差平方和对于β来说是一个二次函数，是一个高维度的抛物面而且开口向上，因此存在着最低值。为了能够画出来，我假设β向量只有2个元素，那么RSS就是一个三维抛物面。它本身及的平面投影示意图如下：



想要求出β向量中的那两个的无偏估计值，只需要对RSS求β的一阶偏导，然后让其为0，接下来就可以求得解析解。但是如果β向量中存在着自变量的共线性，那么上一步计算的最后表达式中（如下）

$$\frac{\partial RSS}{\partial\beta}=0,\ \text{即}$$

$$\frac{\partial\left[\sum_{i=1}^{N}\left(y_i-\beta_0-x_i^T\beta\right)^2\right]}{\partial\beta}$$

$$=\left(x_i^Tx_i\right)\beta-x_i^T\left(y_i-\beta_0\right)=0$$

等价于

$$\beta=\left(x_i^Tx\right)^{-1}x_i^T\left(y_i-\beta_0\right)$$

β有解的充分必要条件rank(XTX)必须是满秩就不成立。或者即便成立，但由于自变量之间高度线性相关，行列式XTX非常接近于0，那么得到的β解就非常不稳定。

这时候，就需要对RSS加入约束条件，或惩罚项。

Elastic net 模型家族是这样给RSS最小化添加约束条件的

$$\min_{\beta_0,\beta}\left[\frac{1}{2N}\sum_{i=1}^{N}\left(y_i-\beta_0-x_i^T\beta\right)^2+\lambda(1-\alpha)\|\beta\|_{L2}^2+\lambda\alpha\|\beta\|_{L1}\right]$$

有些读者更喜欢看代数形式，例如我

$$\min_{\beta_0,\beta}\left[\frac{1}{2N}\sum_{i=1}^{N}\left(y_i-\beta_0-x_i^T\beta\right)^2+\lambda\sum_{j=1}^{p}\left[\frac{1}{2}(1-\alpha)\beta_j^2+\alpha\left|\beta_j\right|\right]\right]$$

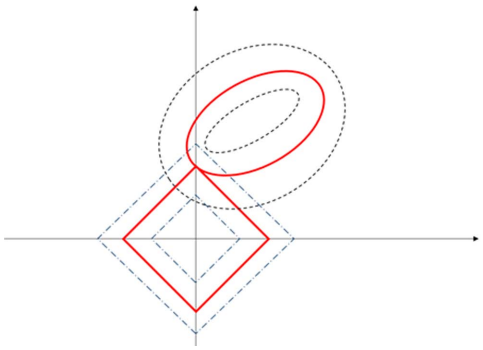
其中n=自变量个数，β_j为β向量的元素

当上式α=0时，最小二乘线性回归就变成了岭回归，此时β向量中所有的元素，也就是模型的系数都保留，但是各个自变量之间通过相互借用影响强度（borrow strength from each other）来保证所有的系数都不为零；

当α=1时，则变成了LASSO回归，此时将会将一些影响力高的自变量保留，影响力不大的则让其系数为0。

这个时候RSS优化使用约束条件，用几何表示就是这样：

还是二维空间的例子，由于|β|中只有2个元素，因此|β|是一个正方形，它和RSS在相同平面上有一个投影交集，就是下图中红色曲线的交点。这个交点对于|β|来说更容易出现在坐标轴上，因此β向量就不必全部取值，只需要取一个就行，另一个就是0。这样就实现了自变变量的选择和过滤。



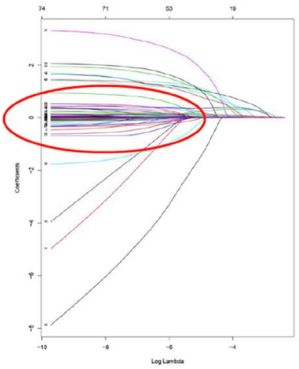
接下来，可以通过设对下面的RSS进行求解极值就可以得到β向量的值。

$$\frac{\partial\left[\frac{1}{2N}\sum_{i=1}^{N}\left(y_i-\beta_0-x_i^T\beta\right)^2+\lambda\sum_{j=1}^{p}\left|\beta_j\right|\right]}{\partial\beta_j}=0$$

由于λ的取值越大，对于整个RSS的极值求解来说就会增加约束条件的难度，因此会被过滤掉更多的自变量。因此，需要借助计算机软件对λ进行逐步迭代试算，找到比较合适的λ值。

在钟院士团队的论文中，提到了R软件的包glmnet，团队通过这个开源的包进行重症病患预测因子的过滤。

通过观察论文中λ-系数轨迹图，我们可以看到，当λ迭代到一定程度，例如log(λ)=-5或-6时，一些不重要的变量的系数全部收敛到接近于0，此时剩下的变量的系数依然有比较显著的非零特性。而刚好这些剩下变量就是之前评分中的那些检查指标（这个结果是利用Logistics回归中的LASSO选择结果，因此剩余的系数个数刚好是10个）。



而当0<α<1时，β向量的结果介于岭回归和LASSO回归之间。

二、对于治病救人而言，模型的使用更应该侧重于可解释。

这一点也是本次论文给我的启示。

其实模型可解释在我日常工作中也经常遇到这样的场景：

辛苦搞了几个模型出来，在平衡预测精度和可解释时，往往很难拿捏。于是乎就选择了“黑盒子”与“可解释”中各挑一个模型，然后加权平均。如果客户也是和我一样是分析师圈子的人，那么他可能就会偏向于“黑盒子”、而若客户是偏业务的人，那么就会更偏向于“可解释”。我还依稀记得客户问我，什么是模型系数的显著性水平、模型的现实意义是啥……这些记忆片段。由于我平时的工作还不涉及到治病救人这么事关人命的大事，因此这种平衡方法使用也就没什么大碍。但是对收治的病人并进行重症的预判，则是一件人命关天的大事儿，来不得半点“讲不清楚，反正就是预测效果好”。另外，我觉得模型选择侧重于可解释，主要还有一个重要原因：

就本论文而言，这个重症评分工具的推广对象，是全社会各级医疗和疾控机构，大家的水平层次都是参差不齐的。当工具的使用对象是只具备基础医疗训练的基层医务人员时，你搞一个难以解释的“黑盒子”，又还要他们懂得设定参数和超参数，这样就会造成基层医务人员的困惑，甚至不知道怎么用，也就难以推广。另外，一些普通老百姓也会看到这个工具，当然有条件自检的话，同样也可以自己使用。如果让普通人去调整这么复杂的模型，这个工具自然就会被敬而远之，起不到造福百姓的作用了。

/ 如何使用数据分析技术进行销售预测 /

作者 /广西CPDA数据分析师 陈虹坚　编辑 / 协会会员处 李苗苗　日期 / 2020-05

在我们的日常工作中，有很多时候，都需要对明年的销售额进行预测，以便进行费用、资金及资源的安排。如何对次年的销售额进行相对合理的预测，相信是市场运营人员及管理者关注的问题。本文通过一个例子，介绍如何使用数据分析技术，对次年的销售额进行预测。

一、业务背景

下表为A公司2014-2017年每个季度的销售额，现需要对该公司2018年的销售额进行预测，以便管理者安排运营资金、资源及进行绩效考核目标设定。

年份	季度	销售额	年份	季度	销售额
2014	1	25131	2016	1	50734
2014	2	25524	2016	2	52437
2014	3	25651	2016	3	48297
2014	4	25516	2016	4	46422
2015	1	34464	2017	1	43096
2015	2	44167	2017	2	47842
2015	3	47009	2017	3	49256
2015	4	49788	2017	4	53262

二、业务分析

经过分析，A公司的产品销售有以下几个特点：

1.销售有淡旺季之分

每年的4-5月、10-11月为销售旺季，除此之外，每年春节、五一、十一等长假也会出现销售高峰；

2.呈现明显的周期变化

该产品每年的每个季度销售特征都非常明显，一季度春节期间销售额达到一个高峰，二季度因国际大环境影响4-5月达到第二次销售高峰，三季度天气炎热为全年销售低谷，四季度随着十一长假及国际大环境影响又迎来一个销售高峰，次年周而复始。

3.因市场及政策因素导致销售的不确定性

该产品有可能会受到国家政策调整、市场大环境及其他不确定因素的影响，导致销售额呈现不规则的波动。

4.长期的趋势

该产品从长期趋势上看，呈现向上增长的态势。

结合上述对A公司产品销售额的分析，我们需要根据该产

品历史的销售数据，结合产品的特点，来预测它未来的销售额。可以采用时间序列预测的季节分解法来对该产品下一年度的销售额进行预测。

三、分析模型

结合上述业务场景，A公司产品销售时间序列主要由四个影响因素构成：长期趋势，用T表示；季节变化，用S表示；周期变化，用C表示；不规则变化，用I表示。

我们设销售额为Y，那么Y即可看成是T,S,C,I的函数，即Y=f(T,S,C,I)，此时只需要确定f的表达式，并且从Y中把T,S,C,I分离出来，那么如果T,S,C,I的变化是已知的，即可对Y进行预测。

根据A公司的产品业务特点，本例时间序列应用乘法模型，即Y=T×S×C×I。

四、预测步骤

1、给变量t赋值；

按照时间的顺序，给现有的历史数据及需要预测的数据进行赋值，如下表所示。表中17-20序列即是需要求出的2018年四个季度的预测数据。

t	年	季度	销售额
1	2014	1	25131
2	2014	2	25524
3	2014	3	25651
4	2014	4	25516
5	2015	1	34464
6	2015	2	44167
7	2015	3	47009
8	2015	4	49788
9	2016	1	50734
10	2016	2	52437
11	2016	3	48297
12	2016	4	46422
13	2017	1	43096
14	2017	2	47842
15	2017	3	49256
16	2017	4	53262
17	2018	1	
18	2018	2	
19	2018	3	
20	2018	4	

2、剔除季节变化；

因为我们的历史数据是按季度呈现的，一年之内变化四次，为了剔除一年之内的波动，需要对数据进行四项移动平均。

四项移动平均后，提出了一年之内的季节变化和不规则变动，即剔除掉了S×I。

t	年	季度	销售额	四项移动平均
1	2014	1	25131	
2	2014	2	25524	
3	2014	3	25651	25455.5
4	2014	4	25516	27788.75
5	2015	1	34464	32449.5
6	2015	2	44167	37789
7	2015	3	47009	43857
8	2015	4	49788	47924.5
9	2016	1	50734	49992
10	2016	2	52437	50314
11	2016	3	48297	49472.5
12	2016	4	46422	47563
13	2017	1	43096	46414.25
14	2017	2	47842	46654
15	2017	3	49256	48364
16	2017	4	53262	
17	2018	1		
18	2018	2		
19	2018	3		
20	2018	4		

3、求出T×C；

我们在上一步骤剔除了S×I后，得出的结果还不是T×C，因为在四项移动平均中，我们已将相邻的4个数据相加取平均得到一个值，表中2014年的四个季度的数据被平均时，它们的平均数应该置于2.5的位置，第二个数应放在3.5的位置，其余数据取平均时也有类似的问题。因此我们需要再做一次居中平均来解决这个问题，如下图所示。完成居中平均后，得到的结果就是T×C。

t	年	季度	销售额	四项移动平均	居中平均（T*C）
1	2014	1	25131		
2	2014	2	25524		
3	2014	3	25651	25455.5	26622.125
4	2014	4	25516	27788.75	30119.125
5	2015	1	34464	32449.5	35119.25
6	2015	2	44167	37789	40823
7	2015	3	47009	43857	45890.75
8	2015	4	49788	47924.5	48958.25
9	2016	1	50734	49992	50153
10	2016	2	52437	50314	49893.25
11	2016	3	48297	49472.5	48517.75
12	2016	4	46422	47563	46988.625
13	2017	1	43096	46414.25	46534.125
14	2017	2	47842	46654	47509
15	2017	3	49256	48364	
16	2017	4	53262		
17	2018	1			
18	2018	2			
19	2018	3			
20	2018	4			

4、求S×I；

因为Y=T×S×C×I，因此S×I=Y/T×C,上一步骤中，我

们已经求出T×C，带入上述公式即可得到S×I，如下表。

t	年	季度	销售额	四项移动平均	居中平均（T*C）	S*I
1	2014	1	25131			
2	2014	2	25524			
3	2014	3	25651	25455.5	26622.125	0.963522
4	2014	4	25516	27788.75	30119.125	0.847169
5	2015	1	34464	32449.5	35119.25	0.981342
6	2015	2	44167	37789	40823	1.081915
7	2015	3	47009	43857	45890.75	1.024368
8	2015	4	49788	47924.5	48958.25	1.016948
9	2016	1	50734	49992	50153	1.011585
10	2016	2	52437	50314	49893.25	1.050984
11	2016	3	48297	49472.5	48517.75	0.99545
12	2016	4	46422	47563	46988.625	0.987941
13	2017	1	43096	46414.25	46534.125	0.926116
14	2017	2	47842	46654	47509	1.007009
15	2017	3	49256	48364		
16	2017	4	53262			
17	2018	1				
18	2018	2				
19	2018	3				
20	2018	4				

5、求T和C；

模型中T表示长期趋势，根据业务背景，A公司的产品销售额从长期趋势上看，是呈上升态势的，因此它是一条反映增长趋势的直线，即T=A+Bt，可以用线性回归的方法求出T，如下图所示。

SUMMARY OUTPUT

回归统计	
Multiple R	0.824782362
R Square	0.680265944
Adjusted R Square	0.657427797
标准误差	6232.373934
观测值	16

方差分析

	df
回归分析	1
残差	14
总计	15

	Coefficients
Intercept	26107.4
t	1844.688235

t	年	季度	销售额	四项移动平均	居中平均（T*C）	S*I	T	C
1	2014	1	25131				27952.09	
2	2014	2	25524				29796.78	
3	2014	3	25651	25455.5	26622.125	0.963522	31641.46	0.841368
4	2014	4	25516	27788.75	30119.125	0.847169	33486.15	0.89945
5	2015	1	34464	32449.5	35119.25	0.981342	35330.84	0.994011
6	2015	2	44167	37789	40823	1.081915	37175.53	1.098115
7	2015	3	47009	43857	45890.75	1.024368	39020.22	1.176076
8	2015	4	49788	47924.5	48958.25	1.016948	40864.91	1.198051
9	2016	1	50734	49992	50153	1.011585	42709.59	1.174279
10	2016	2	52437	50314	49893.25	1.050984	44554.28	1.119831
11	2016	3	48297	49472.5	48517.75	0.99545	46398.97	1.045664
12	2016	4	46422	47563	46988.625	0.987941	48243.66	0.973986
13	2017	1	43096	46414.25	46534.125	0.926116	50088.35	0.929041
14	2017	2	47842	46654	47509	1.007009	51933.04	0.914813
15	2017	3	49256	48364			53777.72	
16	2017	4	53262				55622.41	
17	2018	1					57467.1	1.032444
18	2018	2					59311.79	1.044253
19	2018	3					61156.48	1.021036
20	2018	4					63001.16	1.023829

有了T的值后，因为T×C已知，根据模型Y = T×S×C×I，我们即可求出C的值，如上图所示。由于我们需要求出17-20的C值，则需要根据现有数据，结合业务特征，拟合出一条波动的曲线，得到C在17-20序列的值。

6、求S和I；

已知S×I的值，可以将S和I分别分离出来。S表示季节变化，不同的年份，季节的值应该都是相同的，但我们看下表，同样的年份同样的季度，值并不相同，原因就在于I的影响。为了消除I的影响，我们对不同年份，同一季度的值进行平均。同时，模型中季节因素对Y无影响则四个季度之和应该为4，否则需要进行修正。按等比例原则，修正后的积极指数如下表所示。

S*I	季度				
	1	2	3	4	
2014			0.963521883	0.84716936	
2015	0.981342141	1.081914607	1.024367656	1.01694811	
2016	1.011584551	1.050983851	0.995450119	0.98794123	
2017	0.926116049	1.007009198			
平均数	0.973014247	1.046635885	0.994446552	0.95068624	3.96478292
季节指数	0.98165702	1.055932601	1.003279697	0.95913068	4

7、运用模型预测

我们分别求出了T、S、C的值，根据乘法模型，即可得出2018年每个季度的预测值，计算结果见下表。

t	年	季度	销售额	四项移动平均	居中平均（T*C）	S*I	T	C	S
1	2014	1	25131				27952.09		
2	2014	2	25524				29796.78		
3	2014	3	25651	25456	26622	0.964	31641.46	0.8414	
4	2014	4	25516	27789	30119	0.847	33486.15	0.8995	
5	2015	1	34464	32450	35119	0.981	35330.84	0.994	
6	2015	2	44167	37789	40823	1.082	37175.53	1.0981	
7	2015	3	47009	43857	45891	1.024	39020.22	1.1761	
8	2015	4	49788	47925	48958	1.017	40864.91	1.1981	
9	2016	1	50734	49992	50153	1.012	42709.59	1.1743	
10	2016	2	52437	50314	49893	1.051	44554.28	1.1198	
11	2016	3	48297	49473	48518	0.995	46398.97	1.0457	
12	2016	4	46422	47563	46989	0.988	48243.66	0.974	
13	2017	1	43096	46414	46534	0.926	50088.35	0.929	
14	2017	2	47842	46654	47509	1.007	51933.04	0.9148	
15	2017	3	49256	48364			53777.72		
16	2017	4	53262				55622.41		
17	2018	1	58243				57467.1	1.0324	0.982
18	2018	2	65401				59311.79	1.0443	1.056
19	2018	3	62648				61156.48	1.021	1.003
20	2018	4	61866				63001.16	1.0238	0.959

五、结论与建议

季节分解法的计算思路为时间序列是由时间t和根据时间变化的Y构成，Y由T,S,C,I即长期趋势、季节变化、周期和不规则变动构成，根据乘法模型Y = T×S×C×I，只需要分解T,S,C,I即可求出预测值。同时，任何的数据分析方法都需要与业务特征、行业特点相结合,才能更好得利用数据分析这个工具来解决业务问题。

/ CPDA数据分析师创业沙龙精彩回顾，这是一场创业者的饕餮盛宴 /

来源 /CPDA数据说　编辑 / 协会会员处 李苗苗　日期 / 2020-05

6月5日，我们为全国CPDA数据分析师们带来了一场精彩的主题为《大数据创业蓝海，师兄给你打个样儿》的线上沙龙活动。

本次活动特邀海南智企数据分析师事务所创始人、CPDA数据分析师周召安和大家分享他的创业秘籍，活动吸引了各行业、各领域的近500名CPDA数据分析师。

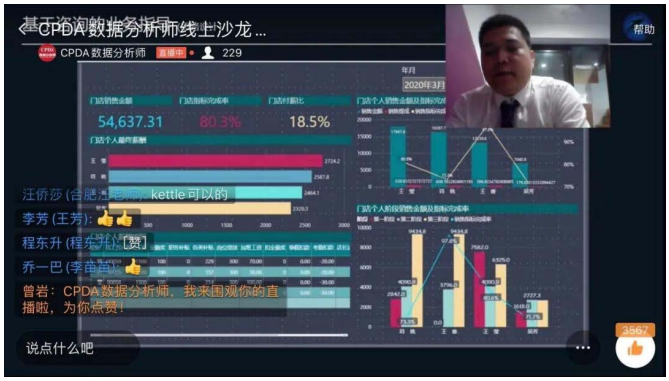
在将近1个半小时的活动中，小伙伴们积极参与到了各环节的互动中，与创业大咖周师兄零距离互动交流、共同探讨创业经验，交流创业心得，分享创业项目，现在就让我们一起来回顾吧！

线上沙龙回顾一：了解数据分析师事务所及案例类别

沙龙活动一开始，周师兄就以海南智企数据分析师事务所为例，为大家介绍了该事务所的三种主要业务类型——基于工具的技术支持、基于技术的产品研发和基于咨询的业务指导，让大家对数据分析师事务所的主营业务有了一个比较清晰的理解；

随着数据复杂性的增加，企业业务人员及决策者对数据分析师的要求也会越来越高，只有真正懂业务的数据分析师，才能分析出什么是对经营业绩最重要的指标，以及需要采取的行动计划有哪些。

线上沙龙回顾二：3个典型案例，带你从方法论到业务实战落地



紧接着，周师兄通过从海南智企数据分析师事务所知识库中抽取的3个真实实战案例：（1）以某家中型医美企业为例的“基于技术的数据产品案例（技术实现）”和（2）以化肥进出口贸易企业为例的“基于工具的支持案例（方法论应用）”（3）以某餐饮企业为例的“基于咨询的业务指导（思路设计）”；帮助CPDA数据分析师们深入洞察了在企业分析时（以自动清洗客户消费行为、客户价值分析、流失概率分析为例）有哪些需要注意的地方，有哪些优化决策的原则和方法，以及如何将所学知识与数据分析实际业务进行完美嫁接。

最后，周师兄就大家关心的个性化问题进行了现场答疑，小伙伴们纷纷表示本次沙龙活动不仅有丰富的理论干货，还有落地的企业经验分享，让CPDA数据分析师们大呼过瘾，收获颇丰!

海南智企数据分析师事务所成立于2018年，事务所主要为广大企业与政府提供数据分析技术服务，包含但不限于数据处理、数据挖掘、数据可视化、数据运用咨询等相关服务。欢迎广大数据分析师和客户朋友与智企数据分析师交流。

海南智企数据分析师事务所负责人周召安 联系电话：18976955759（同微信）

数据场 CODEHALO

广州数据场科技有限公司
CPDA广州授权管理中心

广州数据场科技有限公司拥有超过十年市场运营、国际猎头、企业培训经验的专业团队，与众多中大型企业保持着良好的合作关系。

得益于中国商业联合会数据分析专业委员会的指定授权，开展CPDA数据分析师在广东地区的认证培训工作。肩负起为广东地区培养大数据人才供给的重任。大数据的热度和应用必将形成广州数据场科技有限公司独一无二的庞大社群资源。业务延伸范围必将逐步迅速加深至国内外大数据企业的专业实训落地、大数据职业猎头服务及大数据分析事务所集群！

广州地区科技互联网产业发达，在大数据产业方面，已经形成气候，对数据人才的需求迫切！缺口已达数十万，未来数据专业人才的薪资必将水涨船高，CPDA数据分析师课程内容将数据分析技术与企业运营决策实务结合起来，旨在培养大数据时代能够有效对数据进行综合应用的数据分析专业化、实用型人才，为国家大数据产业发展培养专业人才。

宗旨：用数据说话，做理性决策。

愿景：让数据分析改变每个人的未来，分析引领大数据落地，搭建大数据活动生态圈。

联系方式：王老师 020-39283117 / 13352892978
咨 询 QQ：2693634131（微信同号）
电子邮件：2693634131@qq.com
培训网址：www.gz-cpda.com
办公地址：广州市天河区天河北路183号大都会广场.1401A