— 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
2019年第03期 总第39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 主办

中国商业联合会数据分析专业委员会

祝福祖国七十华诞，为推动大数据战略发展奉献芳华

金秋岁月、丰收的季节，几度风雨几度秋，祖国华诞七十年，七十年的披荆斩棘，七十年的风雨同舟，新中国取得了举世瞩目的成就，民族独立、国家富强、百姓安居乐业是我们共同的中国梦。近几年，大数据发展日新月异，离不开国家对大数据战略的推动，数字基础设施的完善，以及数据资源整合和开放共享的推进……

2014年，国务院《政府工作报告》对大数据发展作出了明确部署；

2015年，国务院印发了《促进大数据发展行动纲要》；

2016年，国家大数据“十三五”规划《大数据产业发展规划（2016——2020）年》出台；

2017年，习近平总书记在十九大报告中提出建设网络强国、数字中国、智慧社会。强调：大数据发展日新月异，我们应该审时度势、精心谋划、超前布局、力争主动，深入了解大数据发展现状和趋势及其对经济社会发展的影响，分析我国大数据发展取得的成绩和存在的问题，推动实施国家大数据战略，加快完善数字基础设施，推进数据资源整合和开放共享，保障数据安全，加快建设数字中国，更好服务我国经济社会发展和人民生活改善；

2018年，习近平总书记提出要以“一带一路”建设等为契机，加强同沿线国家特别是发展中国家在网络基础设施建设、数字经济、网络安全等方面的合作，建设“21世纪数字丝绸之路”；

2019年，在中国国际智能产业博览会上，习近平总书记发来贺信，表明中国对大数据、智能化发展的高度重视。指出“中国愿同国际社会一道，共创智能时代，共享智能成果”，对世界各国携手合作，促进大数据、智能化健康发展必将产生积极而深远的影响。

目前，在我国大数据已经具备了加速发展的良好基础和独特优势的契机与形势下，巨大的潜力和机遇为我们从事大数据的学习、研究以及通过大数据挖掘价值，提供了重要的资源铺垫和政策支持。

协会从2008年成立至今，源于我们对大数据行业的热爱，本着“用数据说话，做理性决策”的信念，初心不忘、深耕不辍：在国家不断加大力度支持大数据行业的背景下，协会不懈奋斗，始终倡导大数据的应用与分析是执业的核心价值；在国内大数据刚刚起步时期，协会致力于Datahoop大数据平台的开发和应用，如今集合了海量算法、国内最大的算法平台，为数据分析师破解算法构建难题；而我们对数据分析师认证及执业管理的科学化以及促进数据分析师事务所组织化、规范化、标准化工作的推进，始终持之以恒用专业和口碑，逐步得到社会的认可与认知；十多年来，更是以为数据分析行业培养复合型人才为己任。我们深信，与大数据同行的学习，才是教育的未来！数据分析作为复合型、交叉型学科，内容涵盖广，学科跨度大，实战要求高，目前市场对于人才的需求远远得不到满足。为了提高高校教师、科研人员以及相关专业从业者对数据分析的理解能力和培养大数据思维，最终促进数据化转型的人才储备，今年8月，协会正式启动了“2019年数据分析讲师公益培养计划”，以行业协会为主导，与全国数据分析师领域优秀教师共创共建，通过专业的能力评估和人才培育咨询，共享数据分析思维、技能、技术和经验，促进数据分析行业的健康发展，推动数据分析人才培养朝更高效、更优质的方向变革。

大数据时代的来临，作为国内数据分析行业的先驱者，我们依然要时刻保持冷静，不忘初心，在数据化变革中传播实用、真实、健康、发展的大数据理念和专业技能，挖掘大数据的实际应用价值，为中国的大数据行业奉献更多的力量。

历史发展，文明繁盛，人类进步，从来离不开思想引领，同时，实践告诉我们，伟大事业都基于创新，创新决定未来。一步一个脚印，一步一个辉煌，在此建国七十周年之际，我代表协会、全体CPDA数据分析师以及数据分析师事务所，向祖国献上深深祝福的同时，也更加坚定信念，不忘初心，不断创新，砥砺前行，不负青春！未来的日子里，协会将继续秉持【专业协会】的发展道路，让我们的CPDA数据分析师及数据分析师事务所，在中国的大数据时代变革中，在国家大数据战略的推动下，体现出数据人更重要的作用与价值，再创新的辉煌！为推动我国大数据事业发展奉献芳华!

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 祝福祖国七十华诞，为推动大数据战略发展奉献芳华

协会动态
04 “全方位育人”——百名数据分析讲师公益培养计划
05 首期“数据分析讲师公益培养计划”圆满落幕

政策导向
07 交通运输部印发《数字交通发展规划纲要》
10 四川省人民政府关于加快推进数字经济发展的指导意见

行业动态
14 数据驱动的决策推动了大数据的采用
16 《大数据时代》用大数据重新定义世界
18 对数据保持敏感将成为未来职场人的分水岭

学数交流
21 大数据助海关精准打击“洋垃圾”走私
23 基于移动云教学平台的学情数据分析实证研究
26 用Python语言做数据分析基本思路和流程
28 运用数据分析方法提升电信存量用户捆绑率的研究

事务所专栏
31 创业项目发展关键，数据的重要性
32 江苏某大型三甲医院数据分析案例
34 奶牛和企业舆情数据分析

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

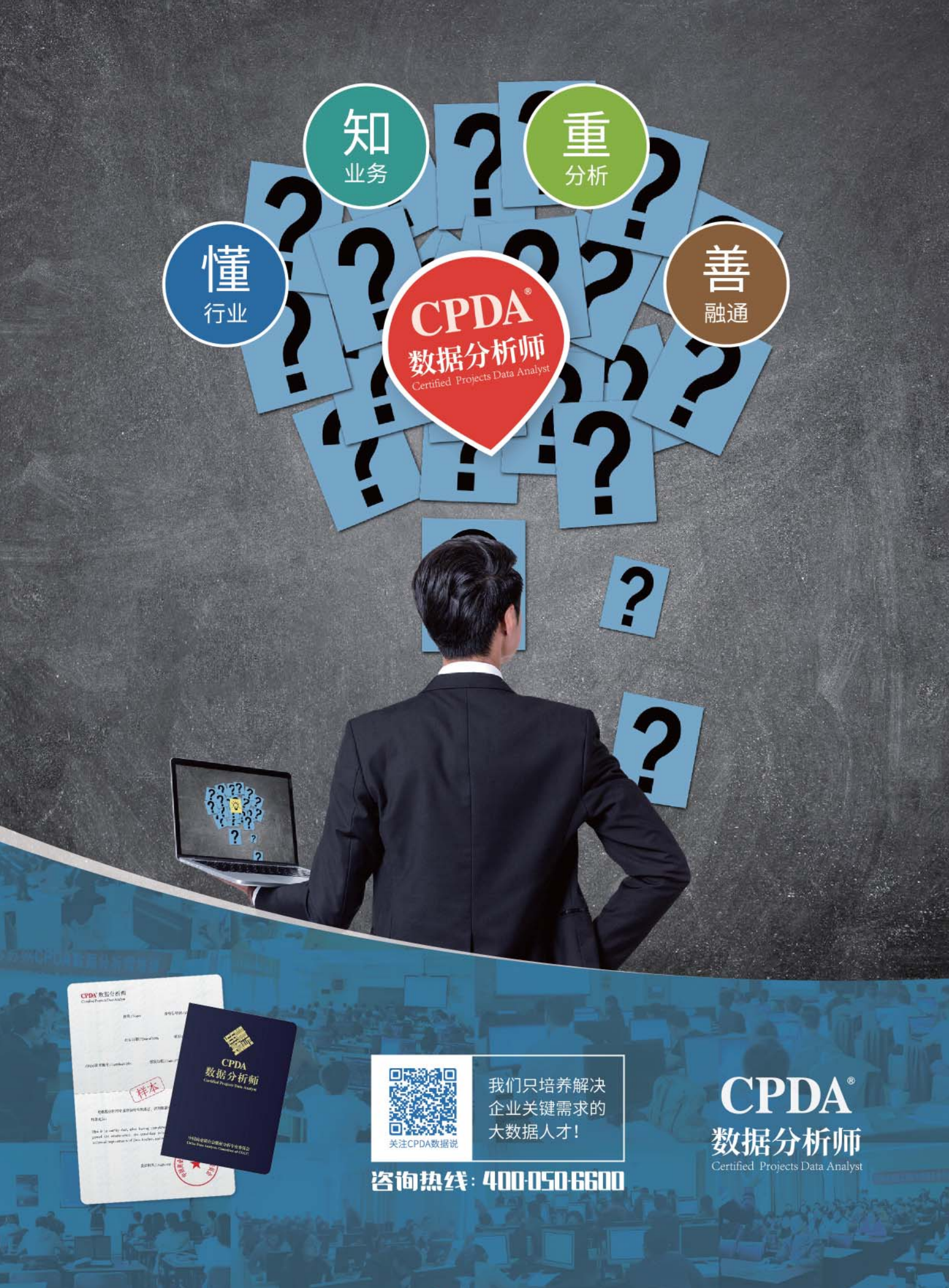
出版时间
2019年9月出版 总第39期

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



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

“全方位育人”——百名数据分析讲师公益培养计划

来源 / CPDA数据分析师网　编辑 / 协会会员处 李苗苗　日期 / 2019-09

数据分析作为复合型、交叉型学科，内容涵盖广，学科跨度大，实战要求高，人才培养面临难题。据国际数据公司IDC预测，随着数据采集、数据存储、数据挖掘、数据分析等数据产业的发展，我国大数据人才缺口数量激增。数联寻英近日发布的首份《大数据人才报告》也显示，目前全国大数据人才只有46万，未来3到5年人才缺口达150万之多，市场需求远远得不到满足。究其原因，还是数据分析师资力量的不足所致。

为了填补数据分析行业的偌大师资缺口，各大院校也争先恐后的开设了大数据专业，而老师们过于“书面化”的授课形式，忽视了实战经验的运用，也成为了各大院校需要迫切解决的问题。

为提高高校教师、科研人员以及相关专业从业者对数据分析的理解和掌握大数据思维，能讲述出学生们爱听的实操课程，最终促进数据化转型人才储备，中国商业联合会数据分析专业委员会正式启动了“2019年数据分析讲师公益培养计划”。以行业协会为主导，邀全国数据分析师资领域优秀教师共创共建，通过专业的能力评估和人才培育咨询，共享数据分析思维、技能、技术和经验，促进数据分析行业的健康发展，推动数据分析人才培养朝更高效、更优质的方向变革。

打破教师职业上升通道，
拒绝照本宣科的教育模式！
为中国数据分析行业的高速发展
提供充足的后备力量↓↓↓

培养计划——协会背书公益扶持

“2019年数据分析讲师公益培养计划”由中国商业联合会数据分析专业委员会主导，现面向社会公开免费培训数据分析讲师百名。面授培训结束后能够具备一定的数据分析讲解能力，能在企业内部展开基础的数据分析培训，或在高校讲授基础的数据分析课程。试讲通过可获得由中国商业联合会数据分析专业委员会颁发的数据分析讲师培训证书，通过公益培训考核和培训的讲师，可以与本次活动支持单位犀数学院签订讲师合作协议，为优秀者推荐单位讲授系列数据分析课程。

面对人群——言传身教于人于己

本计划旨在寻找帮助在数据分析行业内
有强烈意愿成为讲师的人：
研究生及以上学历（条件优秀者可放宽学历要求）
对数据分析有激情，具备丰富业务实战经验的企事业人员
对数据分析知识感兴趣，有意开设和讲授数据分析课程的大中专院校教师
统计学或计算机相关专业，有一定的数学基础知识储备
其他条件：CPDA持证学员优先，研究生及以上学历优先，普通话水平良好，热爱数据分析工作，遵纪守法，品行端正，热爱教育培训事业，具有开拓进取精神和服务意识，具备良好的职业道德。

课程亮点

手把手实操教学让你的执教生涯“稳如泰山”
突破教育边界——打破职业上升瓶颈，由理论型人才培养向实践型人才培养转变。
优化教学思维——针对照本宣科痛点，从知识实践到专业实践，有重点、有重心地进行教学。
打造实战导师——还原一线实战场景，呈现真实解决方案。有疑问现场解答，让你成为真正的“实战导师”。
行业权威背书——报告试讲考核合格，可获得中国商业联合会数据分析专业委员会《数据分析讲师培训证书》，并享有犀数学院讲师资格。

行业权威——专家讲师全方位助力前行

王兴海先生

管理学博士、
应用经济学博士后、高级经济师

曾任上市公司深圳雷柏科技副总裁，日资(丰田电装)、德资及民营等大中型企业工作近20年
擅长：以实际问题为导向，以数据为中心，探索生产和服务系统里的创新，据此发现提升效益的新手段和改善管理的数据化新知识，擅长将量化思维模式融入授课内容。

李妹女士

管理学博士、
中国商业联合数据分析专业委员会特聘顾问

现任：天津财经大学商学院讲师
擅长：产品优化、市场营销、广告效果研究、营销与创新方向，擅长数据问题定位，业务问题数据化，授课思路具备较强逻辑性、课堂引导效果佳。

/ 首期“数据分析讲师公益培养计划”圆满落幕 /

来源 / CPDA数据分析师网　编辑 / 协会会员处 李苗苗　日期 / 2019-09

敬爱的老师
您的谆谆教诲如春风
似瑞雨，永铭我心
九月，是回报的季节
老师，是这季节的主题
教师节到来之际，在这浓情的时光里
让我们向敬爱的老师表达感谢与祝福

当然，教师节对于数据分析讲师公益培养计划的学员们来说，也有着特殊意义：通过这次不同寻常的讲师培养计划，能让他们真正地走上讲台,成为一名合格的数据分析讲师。

2019年9月9日-10日，为期2天的数据分析讲师公益培养计划于北京朝阳的智慧教室正式开班！

面授课程开班的第一天上午，中国商业联合会数据分析专业委员会会长邹东生先生首先发表了精彩致辞，他在致辞中讲到：

近几年伴随着数据分析行业的逐渐热门，执业认证机构已呈现出爆炸式增长，吸引了很多互联网企业纷至沓来投身于“教育事业”，中国商业联合会数据分析专业委员会作为08年成立的，最早投身于数据分析教育领域的领航者，制定了一系列切实可行的人才标准体系，确保每一位持证人才都是符合大数据人才标准的数据分析师。

但截止目前，数据分析行业的人才缺口依旧庞大。目前的教师队伍在数量上也有较大的缺口。为有效推动数据分析教育事业向更科学、更正确的道路上发展，我们不但要关注教师的数量缺口，更要关注教师的质量缺口。

所以在人才培养体系上，中国商业联合会数据分析专业委员会致力于把数据分析人才从“技术型”人才，转变为“经管类”人才，更加强调人才对数据化决策的引导。将原来数据化理解的偏差，逐渐回归到数据应用理念上来。“数据分析讲师公益培养计划”就是特别希望我们很多在校的老师、从事教育工作的人员，能够成为推广行业和认知行业的“践行者”，我们也希望越来越多的老师能跟我们一样，梳理一个正确的、科学的推广方向，把中国商业联合会数据分析专业委员会走了16年的教育理念和知识推广出来，让数据分析行业高质量的教师队伍更加壮大。

由此，我们从没有像今天这样，
对打造更科学、更正确的数据分析讲师
培养的道路充满自信，
对走在大数据与数据分析行业中充满希望，
对担当数据分析育人重任充满信心。

为期两天的公益师资培训，分别从：

1、数据分析人才如何培养与教学（帮助数据分析讲师找准定位）

2、数据分析项目如何操盘，从0到1的策划（帮助数据分析讲师明确自身在每个环节中的角色要求与任务）

3、数据分析项目实战演练（帮助数据分析讲师明确业务实践项目的实现与讲授策略）

4、数据项目报告撰写及探讨（帮助数据分析讲师明确数据项目报告的实现与讲授策略）

5、课程试讲（提升数据分析讲师整体授课能力）几个方面进行了详实的讲解与分享。

在课堂教学中，讲师们精彩的讲授令学员们意犹未尽，分组讨论的环节实现了学员们资源共享、沟通交流、互帮互学、共同提高的目的；

课后学员们就课上讲授的内容热烈讨论、积极互动，各种请教交流不断；课后很多学员也反馈，说老师授课的内容有条理、有重点，刷新了大家对数据分析的整体理解和认知，对学员们既热情又严格，而且老师的授课方式也非常适合大家，希望以后这样的课程多多益善。

/ 交通运输部印发《数字交通发展规划纲要》 /

来源 / 中华人民共和国交通运输部 编辑 / 协会会员处 李苗苗 日期 / 2019-09

据交通运输部官网消息，交通运输部7月25日印发《数字交通发展规划纲要》。规划指出，到2025年，交通运输基础设施和运载装备全要素、全周期的数字化升级迈出新步伐，数字化采集体系和网络化传输体系基本形成。交通运输成为北斗导航的民用主行业，第五代移动通信（5G）等公网和新一代卫星通信系统初步实现行业应用。交通运输大数据应用水平大幅提升，出行信息服务全程覆盖，物流服务平台化和一体化进入新阶段，行业治理和公共服务能力显著提升。交通与汽车、电子、软件、通信、互联网服务等产业深度融合，新业态和新技术应用水平保持世界先进。交通基础设施完成全要素、全周期数字化，天地一体的交通控制网基本形成，按需获取的即时出行服务广泛应用。我国成为数字交通领域国际标准的主要制订者或参与者，数字交通产业整体竞争能力全球领先。

数字交通发展规划纲要

数字交通是数字经济发展的重要领域，是以数据为关键要素和核心驱动，促进物理和虚拟空间的交通运输活动不断融合、交互作用的现代交通运输体系。为贯彻落实党中央、国务院关于促进数字经济发展的决策部署，有力支撑交通强国建设，制定本规划纲要。

一、指导思想

以习近平新时代中国特色社会主义思想为指导，全面贯彻党的十九大和十九届二中、三中全会精神，落实习近平总书记关于加快建设数字中国的重要指示，统筹推进“五位一体”总体布局，协调推进“四个全面”战略布局，按照“巩固、增强、提升、畅通”八字方针，抓住新一轮科技革命和产业变革的机遇，坚持推动高质量发展，坚持以人民为中心，坚持以创新为第一动力，促进先进信息技术与交通运输深度融合，以“数据链”为主线，构建数字化的采集体系、网络化的传输体系和智能化的应用体系，加快交通运输信息化向数字化、网络化、智能化发展，为交通强国建设提供支撑。

二、基本原则

创新引领，数据赋能。以数据为关键要素，赋能交通运输及关联产业，推动模式、业态、产品、服务等联动创新，提升出行和物流服务品质，让数字红利惠及人民，增强人民获得感。

共建共享，融合发展。充分发挥统筹规划、协同推进的制度优势，推动政企、行业、部省间协同发力。发挥市场主体作用，科学配置各类资源要素，构建跨界融合、共创共享的数字交通产业生态。以数据链促进多种运输方式高效衔接，促进

政企间数据双向转化运用。

防范风险，保障安全。兼顾创新发展和安全发展，防范化解数字化转型带来的信息安全风险，提升网络安全和数据安全保障能力，保障公共安全和国家利益。

勇于探索，试点先行。坚持世界眼光、国际标准、中国特色，以开放包容的态度，适应技术发展趋势，以试点为重要手段，汇聚技术、智力、产业等资源，通过典型引路，逐步形成数字交通发展的“中国经验”和“中国方案”。

三、发展目标

到2025年，交通运输基础设施和运载装备全要素、全周期的数字化升级迈出新步伐，数字化采集体系和网络化传输体系基本形成。交通运输成为北斗导航的民用主行业，第五代移动通信（5G）等公网和新一代卫星通信系统初步实现行业应用。交通运输大数据应用水平大幅提升，出行信息服务全程覆盖，物流服务平台化和一体化进入新阶段，行业治理和公共服务能力显著提升。交通与汽车、电子、软件、通信、互联网服务等产业深度融合，新业态和新技术应用水平保持世界先进。

到2035年，交通基础设施完成全要素、全周期数字化，天地一体的交通控制网基本形成，按需获取的即时出行服务广泛应用。我国成为数字交通领域国际标准的主要制订者或参与者，数字交通产业整体竞争能力全球领先。

四、构建数字化的采集体系

（一）推动交通基础设施全要素、全周期数字化。

推动交通基础设施规划、设计、建造、养护、运行管理等全要素、全周期数字化。构建覆盖全国的高精度交通地理信息平台，完善交通工程等要素信息，实现对物理设施的三维数字化呈现，支撑全天候复杂交通场景下自动驾驶、大件运输等专业导航应用。针对重大交通基础设施工程，实现基础设施全生命周期健康性能监测，推广应用基于物联网的工程质量控制技术。

（二）布局重要节点的全方位交通感知网络。

推动铁路、公路、水路领域的重点路段、航段，以及隧道、桥梁、互通枢纽、船闸等重要节点的交通感知网络覆盖。推动交通感知网络与交通基础设施同步规划建设，深化高速公路ETC门架等路侧智能终端应用，建立云端互联的感知网络，让“哑设施”具备多维监测、智能网联、精准管控、协同服务能力。注重众包、手机信令等社会数据融合应用。构建载运工具、基础设施、通行环境互联的交通控制网基础云平台。加快北斗导航在自由流收费、自动驾驶、车路协同、海上搜救、港口自动化作业和集疏运调度等领域应用。

（三）推动载运工具、作业装备智能化。

鼓励具备多维感知、高精度定位、智能网联功能的终端设备应用，提升载运工具远程监测、故障诊断、风险预警、优化控制等能力。推动自动驾驶与车路协同技术研发，开展专用测试场地建设。鼓励物流园区、港口、铁路和机场货运站广泛应用物联网、自动驾驶等技术，推广自动化立体仓库、引导运输车（AGV）、智能输送分拣和装卸设备的规模应用。推动自动驾驶船舶、自动化码头和堆场发展，加强港航物流与上下游

企业信息共享和业务协同。

五、构建网络化的传输体系

推动交通运输基础设施与信息基础设施一体化建设，促进交通专网与“天网”“公网”深度融合，推进车联网、5G、卫星通信信息网络等部署应用，完善全国高速公路通信信息网络，形成多网融合的交通信息通信网络，提供广覆盖、低时延、高可靠、大带宽的网络通信服务。

六、构建智能化的应用体系

（一）打造数字化出行助手。

促进交通、旅游等各类信息充分开放共享，融合发展。鼓励平台型企业深化多源数据融合，整合线上和线下资源，鼓励各类交通运输客票系统充分开放接入，打造数字化出行助手，为旅客提供“门到门”的全程出行定制服务。倡导“出行即服务（MaaS）”理念，以数据衔接出行需求与服务资源，使出行成为一种按需获取的即时服务，让出行更简单。打造旅客出行与公务商务、购物消费、休闲娱乐相互渗透的“智能移动空间”，带来全新出行体验。推动“互联网+”便捷交通发展，鼓励和规范发展定制公交、智能停车、智能公交、汽车维修、网络预约出租车、互联网租赁自行车、小微型客车分时租赁等城市出行服务新业态。

（二）推动物流全程数字化。

大力发展“互联网+”高效物流新模式、新业态，加快实现物流活动全过程的数字化，推进铁路、公路、水路等货运单证电子化和共享互认，提供全程可监测、可追溯的“一站式”物流服务。鼓励各类企业加快物流信息平台差异化发展，推进城市物流配送全链条信息共享，完善农村物流末端信息网络。依托各类信息平台，加强各部门物流相关管理信息互认，构建综合交通运输物流数据资源开放共享机制。

（三）推动行业治理现代化。

完善国家综合交通运输信息平台，提高决策支持、安全应急、指挥调度、监管执法、政务服务、节能环保等领域的大数据运用水平，实现精确分析、精准管控、精细管理和精心服务。完善资源目录与信息资源管理体系，实现行业信息资源的汇聚融合，提升信息资源共享交换和开放服务能力。建立大数据支撑的决策与规划体系，推动部门间、政企间多源数据融合，提升交通运输决策分析水平。采用数据化、全景式展现方式，提升综合交通运输运行监测预警、舆情监测、安全风险分析研判、调度指挥、节能环保在线监测等支撑能力。进一步推进交通运输领域“互联网+政务服务”，实现政务服务同一事项、同一标准、同一编码。延长网上办事链条，推动政务服务向“两微一端”等延伸拓展。加快完善运政、路政、海事等政务信息系统，推进交通运输综合执法、治超联网等系统建设，提高执法装备智能化水平，推进在线识别和非现场执法。

七、培育产业生态体系

聚焦基础设施和载运工具数字化的关键环节与核心技术，鼓励优势企业整合电子、软件、通信、卫星、装备制造、信息服务等领域资源，构建强强联合、优势互补、高效适配的协同创新体系。加强测试、检测、认证综合能力建设，促进新技术成果转化。加快北斗导航、卫星通信、高分辨率对地观测等技术行业应用。鼓励建立协同创新产业联盟，积极开展产业化应用示范，促进各类主体合作，打造具有国际竞争力的产业生态体系。

八、健全网络和数据安全体系

贯彻落实《中华人民共和国网络安全法》以及国家关于

数据安全的要求，落实各级交通运输管理部门及相关机构的网络安全职责，健全信息通报、监测预警、应急处置、预案管理等工作机制，建立专家库。落实网络安全等级保护制度，确保各级安全防护合规达标。加强网络安全与信息系统同步建设，提高交通运输关键信息基础设施和重要信息系统的网络安全防护能力。推进重要信息系统国产密码应用、重要软硬件设备国产化应用。加强对交通数据全生命周期的管控，保护国家秘密、商业秘密和个人隐私。完善适应新技术发展的行业网络安全标准。

九、完善标准体系

加快完善面向数字交通应用的交通基础设施工程建设标准，推动信息基础设施与交通基础设施同步规划、同步设计、同步建设、同步运维。按照交通运输信息化标准体系，持续完善相关标准。加快自动驾驶国家及行业标准体系建设，完善生产制造、测试评价、网络安全、数据共享、运行使用等标准。推动建立跨行业、跨领域、跨部门数字交通标准协同发展机制，发挥企业在标准研究方面的作用，积极参与国际标准制修订的协调、交流与合作。

十、完善支撑保障体系

（一）营造发展环境。

探索有利于数字交通创新发展的行业监管模式，推动完善相关法律法规体系，在规范发展、安全发展的前提下，营造有利于创新发展的政策环境。积极推进企业间、政企间的数据资源融合应用，鼓励市场主体提供丰富的数字交通服务，激发创新活力和潜在价值，提供更好的服务体验。

（二）多渠道筹措资金。

发挥中央财政资金的引导作用，加大对创新试点支持力度，强化过程指导，加强绩效评估。各级交通运输主管部门积极争取财政性资金、专项资金等支持数字交通建设。探索政府和社会资本合作模式。

（三）促进创新应用。

坚持试点先行，带动全面发展，围绕“一带一路”建设、京津冀协同发展、长江经济带发展、粤港澳大湾区发展、长江三角洲区域一体化发展等国家重要区域发展战略，开展区域性数字交通综合试点。系统开展数字交通发展配套政策研究。在保障国家安全、维护国家利益的前提下，促进资源跨境流动，实现引资、引智、引技相结合。

（四）加强人才支持。

建立跨领域、多层次人才培养体系，提升行业数字化思维和应用能力。建立多层次专家库，发挥高端智库、院校等机构的智力支持作用，鼓励建立产学研金的对接平台。

/ 四川省人民政府关于加快推进数字经济发展的指导意见 /

来源 / 四川省人民政府　编辑 / 协会会员处 李苗苗　日期 / 2019-09

各市(州)人民政府，省政府各部门、各直属机构，有关单位：

为深入贯彻落实党中央国务院关于加快发展数字经济的战略部署，全面落实省委省政府关于加快构建现代产业体系的意见要求，加快建设网络强省、数字四川、智慧社会，形成具有较强核心竞争力的数字经济生态体系，结合我省实际，现提出如下指导意见。

一、总体要求

(一)指导思想。以习近平新时代中国特色社会主义思想为指导，按照高质量发展总体要求，围绕“一干多支、五区协同”“四向拓展、全域开放”战略部署，以数据为关键要素，以“创新驱动、融合发展，市场主导、重点突破，开放共享、安全规范”为发展原则，以“数字产业化、产业数字化、数字化治理”为发展主线，加快推进数字经济发展，为推动治蜀兴川再上新台阶提供强力支撑。

(二)发展目标。到2022年，全省数字经济总量超2万亿元，成为创新驱动发展的重要力量。数字经济核心产业规模不断壮大，电子信息产业在全省经济高质量发展中核心支撑作用更加凸显。数字经济与实体经济融合发展水平显著提高，两化融合总指数超90。提升数字化治理能力，为数字经济活跃发展营造良好环境。

二、加快发展数字经济核心产业

(三)做大做强大数据产业。加快推进大数据产业集聚区和产业园建设，打造“成德绵眉泸雅”大数据产业集聚区，建设3—5个大数据产业基地。依托重点园区发展大数据流通交易、技术服务、科研“双创”等公共平台。加强数据采集、存储管理、挖掘分析、安全保护等领域关键技术攻关，形成一批自主创新、技术先进、满足重大应用需求的产品、解决方案和

服务应用。打造大数据应用场景，推进政务服务、普惠民生、公共服务、产业创新领域大数据应用，加快数字经济与实体经济融合发展。努力建成全国重要的大数据生态建设高地、大数据研发创新高地、大数据示范应用高地和大数据人才集聚高地。

(四)加快发展人工智能产业。积极建设创新平台，加强人工智能领域基础理论研究与关键共性技术攻关，培育智能机器人、无人机等人工智能重点产品和人工智能龙头企业。加快建设天府新区人工智能产业聚集区、智能制造产业园，积极申建国家级人工智能创新试验区和产业示范园区，加快培育建设人工智能产业创新集群。打造人工智能深度应用场景，推进中德智能网联汽车等区域示范，推动人工智能技术在社会各领域大规模应用。

(五)促进5G产业突破发展。大力建设“天府无线通信谷”、中国移动(成都)产业技术研究院、中国电信云锦天府5G应用产业园、中国联通5G创新中心等产业载体和创新平台。打造5G产业领域重点产品，构建5G完整产业链。开展成都天府国际机场5G网络商用示范，推动“一杆多用”(见尾页名词解释)试点和窄带物联网(NB—IoT)应用示范，率先在主要城市主城区实现5G规模商用。实施5G在车联网、智慧医疗、智慧物流等领域示范工程。

(六)大力发展超高清视频产业。加强超高清产业整体布局，建设成都影视硅谷、超高清视频(四川)制作技术协同中心，打造国家超高清视听创新基地。积极推动4K/8K产业的图像传感器、核心芯片、显示面板、终端设备等核心产品技术研发和关键设备产业化。开展5G VR 4K/8K和5G 4K AI视频监控示范工程，推动超高清视频产业与垂直行业深度融合，争创国家超高清视频产业基地。

(七)巩固发展电子信息基础产业。聚焦“一芯一屏”，着力推进“设计—制造—封装测试—材料设备—信息服务”产业链一体化发展。重点发展集成电路设计、制造、封测，配套发展集成电路材料和设备业，打造全国领先、中西部地区最大的集成电路产业基地。聚焦柔性显示、透明显示两个方向，打造国际一流、国内最大的新型显示产业基地。积极落实软件企业税收优惠政策，重点支持软件产品、软件服务和嵌入式系统三大产业方向，打造世界知名、国内一流的软件和信息服务业基地。加强密码算法、核心芯片、量子加密、网络安全、数据安全等方向技术攻关，建设省工业信息安全创新中心，打造中西部领先、国内一流的国家级信息安全产业基地。

(八)大力发展数字文创产业。发挥我省特色自然资源、民族资源、文化资源优势，建设全国重要的数字文创中心。建设天府文创城、成都游戏动漫基地、区域数字出版基地、四川传统文化影视内容基地、四川电竞产业基地。支持互联网龙头企业建设分发平台，支持新型主流媒体和龙头企业建立互联网传播平台。大力发展数字媒体、数字出版、3D动漫等数字内容供给。开展大数据和虚拟/增强现实(VR/AR)技术深度融合应用。

三、加快产业数字化转型

(九)创新发展工业互联网。支持五大支柱产业龙头企业建设一批省级工业互联网平台，培育国家级跨区域跨领域工业互联网平台。支持成都建设工业互联网标识解析核心节点，推动工业互联网标识解析体系在典型行业深化应用。实施“万家企业上云”行动，新增上云企业10000家，打造上云示范企业

100家。开展工业互联网集成创新应用试点示范，形成一批面向中小企业的典型应用。实施制造企业互联网“双创”平台建设工程，培育一批支持制造业发展的“双创”示范基地。建设智慧产业园区，搭建四川省产业园区云平台，打造全省产业园区数据库，建设全省统一的园区管理和监督体系。

(十)深入实施智能制造工程。开展智能制造试点示范，推动企业数字化、网络化、智能化发展，培育一批智能制造系统集成商。在航空航天、电子信息、装备制造、汽车制造等重点领域，实施“设备换芯”“生产换线”“机器换工”，建设1000家以上智能工厂/数字化车间，培育100家以上“互联网协同制造”示范企业。

(十一)加快农业农村数字化进程。完善农业农村领域统计监测、预警防控、质量安全、综合服务等信息系统建设，推进农业大数据开放共享。开展全省数字农业示范工程建设，推进现代农业园区数字农业工程。推动农业电商经营体系建设，重点打造“川”字号农产品网上展示平台。推动信息进村入户与基层农技推广体系、农业信息服务体系融合，就近为农民和新型农业经营主体提供培训体验服务。

(十二)推进服务业数字化升级。大力培育网络体验、智能零售、共享经济、平台经济等新模式新业态。实施“领军企业行动”，重点培育食品、农产品、文旅等领域垂直电商供应链平台。发展线下线上结合的跨界业务融合模式，开展餐饮、零售、家政等智慧服务新场景。推动数字商业街区打造和基地建设，创建智慧社区服务示范中心。加快发展跨境电子商务，加快服务贸易数字化进程，推动四川产品和服务“走出去”。

四、加快推进数字化治理

(十三)强化数字政府基础支撑能力。加强电子政务内外网、政务云平台等政府信息化基础设施建设，完善省市县乡村五级互联互通的基础网络体系建设。推动全省政务数据、公共数据、社会数据汇聚融合，为各地各部门(单位)管理、服务、决策提供数据支撑。支持城市公共设施、建筑、电网等领域的物联网应用和智能化改造，推进数字城管与智慧社区融合发展。

(十四)提升政府数字化监管水平。完善省、市两级互联互通的政务信息资源共享交换体系，加快各地各部门(单位)整合共享样板建设，强化数字在政务、市场监管、生态环保、食品安全监管、公共区域监测监控、公共安全等领域的应用。建设“互联网监管”平台，通过大数据提升事中事后监管规范化、精准化和智能化水平。

(十五)提高政府数字化服务水平。加强全省一体化政务服务平台建设，全面实现“一网通办”。依托“12345”政务服务热线平台，健全全省统一的政务服务热线办理机制。推动“最多跑一次”改革向基层、老少边穷地区延伸，充分利用社会第三方拓展办事渠道，实现公共服务“就近办”。推动社会保障卡、居民健康卡、金融IC卡等深度融合应用;加快推进智慧法院、智慧检务和智慧司法建设。

五、深化智慧社会建设

(十六)大力推进智慧教育。深化“三通两平台”建设，整合资源，形成覆盖全省、互联互通的数字教育资源服务体系。鼓励数字校园、智慧校园、未来学校等新模式，开展“省级智慧教育示范区”和“四川省智慧教育学校”创建活动，引导优质教育资源向老少边穷地区覆盖。推动各类学校对接线上线下，开发数字教育资源，开展个性化教育和精准化教学管理，提供多样化、精准化、个性化的学习支持服务。

(十七)积极发展智慧医疗。建设“国家级健康医疗大数据应用中心”。完善互联互通的省、市、县三级全民健康信息平台。完善各类人口和健康相关数据库，实现全民健康信息共享应用和业务协同。发展个性化医疗，建立远程医疗应用体系，开展高质高效的“互联网医疗”健康服务。创新“互联网 居家社区养老”模式，打造一批智慧微型养老院和智慧养老社区。

(十八)加快发展智慧文旅。推进数字化景区、智慧旅游城市建设。打造“智游天府”文旅公共服务平台。构建省、市、县、乡(镇)、企业等智慧文旅和公共服务的一体化体系，建设一批智慧景区、智慧酒店、智慧文旅小镇、数字图书馆、数字文化馆、数字博物馆。加快省文旅大数据中心建设。打造一批文化和旅游数字经济产业园。结合影视新媒体推动文旅融合。探索推动自然景观与虚拟现实技术的融合，提高旅游数字化管理、精准营销和服务智能化水平。

(十九)发展智慧交通物流。大力推进交通设施数字化改造，建设综合交通信息枢纽，搭建综合交通出行信息服务平台;开展平安智慧高速公路建设和智慧普通公路建设试点。完善交通旅游信息采集体系，提供“运游一体化”信息服务产品。开展全省物流大数据中心建设。加快企业物流信息系统建设，实现物流信息全程可追踪。抓好智能仓储系统、冷链物流中心等建设，积极打造无人机配送示范区域。

(二十)积极发展智慧金融。打造天府数字金融产业聚集区。逐步建立金融与政府、其他行业领域的数据资源共享交换机制，加强金融数据的交换共享和协同流动，为金融大数据分析与智能决策提供支撑。做好企业信用信息归集和整理工作，与省社会信用信息平台相衔接，建设数字信用平台。

六、保障措施

(二十一)完善信息基础设施支撑体系。建成高速、移动、安全、泛在的新一代信息基础设施，优化互联网骨干网络架构、扩容省际出口带宽，提高网间流量疏导能力和互通效率。全面推进IPv6规模部署和升级改造。增强窄带物联网(NB—IoT)接入支撑能力。加快“光网四川”“无线四川”“高清四川”建设，实现高速光纤网及4G网络城乡全面覆盖，支持通信企业开展5G网络规模组网及5G业务商业化应用等技术攻

关，持续推进“宽带乡村”工程建设和电信普遍服务试点。利用我省清洁能源优势，争取国家级数据服务中心、云计算中心、超算中心等落户四川。引导数据中心向大规模、一体化、绿色化、智能化方向发展，降低数据中心能耗;提高数据中心数据平均上架率。

(二十二)推进产业创新能力建设。支持工业云制造(四川)创新中心、工业信息安全创新中心、工业大数据创新中心、工业设计研究院、软件开发云等重大创新载体建设。围绕数字经济重点行业领域发展需求，以行业数字化共性关键技术研发为重点，鼓励行业龙头企业、知名高校院所、行业研究机构等在川建设产业创新中心、工程研究中心、重点实验室、企业技术中心等创新平台，开展关键共享技术联合攻关，构建多层次自主创新体系，提高技术研发与创新能力。

(二十三)营造数字经济发展生态。争创国家数字经济创新发展试验区，打造一批数字经济示范城市。因地制宜规划发展各具特色的数字经济发展集群。继续办好世界工业互联网大会、中国国际软件合作洽谈会、中国大数据应用大会、中小微企业云服务大会等品牌会议，打造立足西部、面向全球的数字经济品牌。加大对企业品牌建设的扶持和激励力度，鼓励四川重点企业实施品牌推广战略，着力塑造一批具有较强市场影响力和竞争力的品牌企业、产品和服务。

(二十四)探索数据资源开放与保护。加快数据开放网站建设，推动政府数据向社会开放，鼓励企业和公众发掘利用开放数据资源，不断释放数据经济价值和社会效应。加大对技术专利、数字版权、数字内容产品、个人隐私等的保护力度。研究保障大数据流通交易的政策措施，鼓励企业对脱敏后的数据进行市场交易。

(二十五)健全安全保障体系。针对数据采集、存储、传输、共享和应用过程中的安全问题，加强云平台安全管理，加强网络空间实体身份管理。完善数字安全事件应急预案，健全网络安全风险预警、情报共享、研判处置和应急协调机制。建立关键信息基础设施安全保障体系，提升对关键信息基础设施网络安全突发事件的应急处置能力。增强网络信息安全防护能力，落实网络安全责任制，建立网络信息安全防线，加强重要网络基础设施的安全防护。

(二十六)加强组织领导。发挥省推进数字经济发展领导小组作用，统筹推进全省数字经济发展工作，研究数字经济发展战略、总体规划和政策措施。成立四川省推进数字经济发展专家咨询委员会和四川省数字经济研究院，开展重大前沿问题研究，为重要应用项目及工程实施提供决策咨询。发起设立数字经济发展基金。

(二十七)加大要素保障。对满足布局导向、能效值要求的数据中心和数字化高载能产业项目，实行支持性电价政策。加大金融扶持力度，创新金融支持方式。积极引进数字技术领军人才，建立数字技术及应用人才教育体系，构建企业数字经济人才培训平台。

(二十八)建立统计指标体系和考评机制。完善《四川省数字经济核心产业统计分类目录》，探索开展针对数字经济新领域、新业态和新模式的专项统计研究，明确统计口径，探索数字经济增加值测算方法。建立省级数字经济核心指标的定期发布机制和动态监测制度。建立市州数字经济发展评估体系。

/ 数据驱动的决策推动了大数据的采用 /

来源 / CPDA数据分析师网 编辑 / 协会会员处 李苗苗 日期 / 2019-09

大数据被认为是数据管理和分析的下一个巨大变革。世界各地的许多企业在运营中都使用了大数据技术来帮助他们分析持续生成的数据。

大数据技术的应用显示出许多前景，在终端应用行业中非常受欢迎。随着大数据的不断普及，以及与人工智能和云计算的集成变得更加精简，预计未来还会有进一步的增长。到2022年，全球大数据技术和服务市场的估值将超过1185.2亿美元，从2015年到2022年，其年复合增长率将达到26.0%。

数据驱动决策继续推动大数据技术的应用

多年来，企业做出关键业务决策的方式发生了重大转变。传统的智能和假设已经让位于数据驱动、基于事实的决策，这有助于采用大数据解决方案。

这种现状的改变，已经成为各终端行业越来越多地采用大数据技术和服务的关键因素之一。随着越来越多的企业认识到大数据在决策中的优势，大数据技术和服务的采用很有可能在短期和长期内稳步增长。

大数据分析和信息也帮助许多企业克服了与敏捷性和利益相关者授权相关的挑战。在协调和权力下放之间找到难以捉摸的平衡，企业一直面临着艰巨的任务。

传统上，企业在做出重大决策之前，都要根据每个人的观点。然而，在竞争异常激烈的环境中，这也带来了放慢决策过程的风险。

RACI框架已被企业引用，以减少在选择决策正确权限时的模糊性，随着数据访问变得更容易导航，使整个决策过程成为一个无缝的事务。

大数据技术与传统商业智能的融合——未来之路?

将大数据技术和服务与传统商业智能相结合，被视为企业专注于基于事实的快速决策和改善客户体验的前进之路。

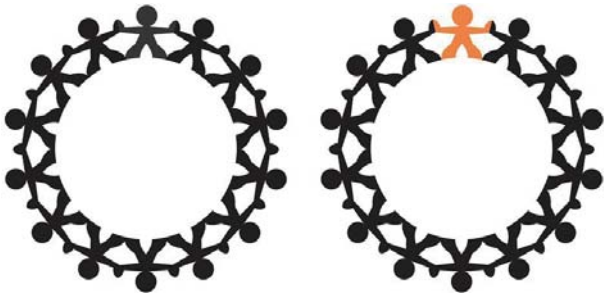
商业智能已成为企业更密切地了解目标受众的可靠工具;然而，高周转时间仍然是一个障碍。大数据的融合在一定程度上缓解了这一挑战，进而推动了终端用户的采纳。未来，大数据和商业智能很可能会高度交织在一起。

银行、金融服务和保险(BFSI)行业继续处于采用的最前沿

虽然大数据技术和服务的应用已经广泛，但金融和保险行业自大数据诞生之初就一直走在行业的前列。这些行业每天产生的大量数据使得采用全面的数据监控、收集和分析解决方案成为必要。

BFSI部门面临的一些关键挑战包括无组织的数据、欺诈识别和运营效率低下。大数据技术和服务的使用在很大程度上缓解了这些挑战。

根据目前的估计，到2026年，采用大数据技术和服务产生的收入可能超过330亿美元。



将大数据技术和服务纳入医疗领域正在取得进展

大数据在医疗保健行业具有巨大的潜力，支持者鼓吹大数据带来的好处包括流行病预测和降低治疗成本。虽然电子健康记录(EHR)在医疗保健领域已经成为一种主流，但其有效性仅限于患者的病史。另一方面，大数据有望提供全面、全面的数据分析，帮助医疗服务提供商管理海量数据。

通过纳入大数据技术和服务提供的洞见，可以帮助医疗服务提供商提高盈利能力，同时改善人们获得的医疗服务。

治理中的大数据:帮助决策者做出更好的决策

除了私营部门越来越多地采用大数据技术和服务外，大数据技术和服务也被纳入治理和行政管理。世界各国政府都面临着一项艰巨的任务，那就是整理有关数亿人的可变数据。收集数据及其组织需要数亿美元的政府支出。

虽然大数据并不能完全取代所有国家的手工和物理流程，但它与传统的数据采集实践相结合，可以帮助实现无缝、快速的数据采集。

总体而言，大数据技术和服务的前景似乎是光明的，它们在一系列最终用途行业的迅速采用，可以进一步促进全球市场的增长。

/ 《大数据时代》用大数据重新定义世界 /

来源 / CPDA数据分析师网　编辑 / 协会会员处 李苗苗　日期 / 2019-09

“改变”是纪录片《大数据时代》在创作中的“关键词”。节目中所讲述的故事都是关于大数据技术对我们的生产、生活所带来的改变。

作为一个时代的到来，似乎总能找到几个标志性的事件，而大数据时代却有所不同，它到来得如此之迅猛，是你我始料未及的，仿佛一夜之间，它已渗透到我们生活中的每个角落。

《大数据时代》选取了当下最为关心和关注的民生热点，大数据不但能够帮助教练挑选优秀的足球运动员，还能为居家养老的老人提供安全保障，无论是医疗或是空气治理，还是最为传统的农业生产，处处都能见到它的身影。

数据，自古有之。如今，那些散落在我们日常生活中的数据正在汇聚，伴随着新一代信息技术的发展，这些数据正在为我们的生活和生产方式，带来深刻的改变。

对于体育教练来说，如何选拔优秀的球员，一直是个难题。在过去教练只能靠经验来选拔，往往会造成部分有潜质的小球员错失了继续训练的机会。

如今，大数据技术在校园的应用就能够改变这一现状，通过对小球员的数据采集和分析，可以科学地选拔出优秀的球员，有效提高球员的训练水平和战术水平。

作为国际化大都市的上海，老龄化形势严峻。静安区已率先进入了深度老龄化社会，全区老龄化人口已经达到了十万以上，其中三个人当中就有一位是老年人。宁绍军是这个平台的搭建者之一，这个平台的首要目的是预防老人在家中发生意外。

三年前，宁绍军就开始参与到这个大数据平台的搭建中，然而通过门禁、烟感、红外、床垫等传感器收集的数据，如何能够帮助社区更好地服务居家养老的老人？是否能够准确预测老人在家中发生意外？又是否能够判断出老人的身体健康状况？

在中国，做一万个小时的手术才能磨练出一位优秀的外科医生，而在一个县级市的基层医院要想达到这个目标，需要二十年。

如今，专家医师的技术正在被数据记录下来。未来，当机器学习到更多专家医师的经验后，机器将会变得越来越聪明，由此也能够辅助更多基层的医生完成高难度的手术。

随着中国环保事业的不断发展，环保执法工作也逐渐步入科学化。

李雪峰，包头市环境监察支队的大队长，排查偷排漏排的污染企业是他的工作日常。因为他所负责的辖区面积较大，如何精准定位污染源成了一个难题。

内蒙古自治区遍布自治区 12 个市盟的 146 个监测站点，如同组成了一张覆盖内蒙古重点区域的监测网，在未来通过设立更多的微型空气监测站，通过风向等数据分析，可以精准地缩小疑似污染源的区域，提高执法效率。

千百年来靠天吃饭的传统农业模式正在悄然改变，无论是育种还是作物生长的全周期，或是收获的时间节点，大数据都能为其提供帮助。

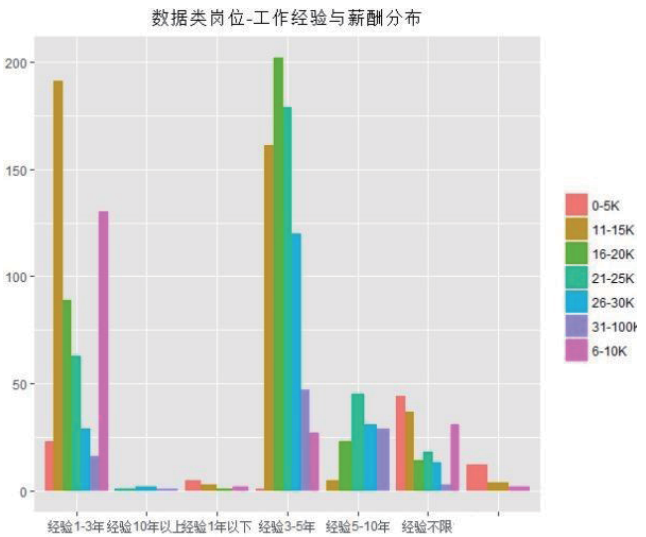
如今的人们，通过采集气象、土壤湿度、稻叶温度等数据，可以有效分析出稻瘟病的发病时间，精准为农民提供合理打药时间进行预防。

新一代信息技术的发展，缩短了世界的距离，将越来越多领域的产品流通变为数据流通，将生产演变成服务，将工业劳动演变成信息劳动。

人类正在迎来大数据新航海时代。

数字中国，序幕已开。

序幕已开，数据分析将改变世界



数据驱动发展的新时代。人类通过科学解决问题，而科学的落脚点就是数据。但不管是大数据还是小数据，其本身并没有含义，只有不断从数据中进行分析洞察，将数据转化为信息和知识，并用来解决业务中的实际问题，才能真正创造出价值。就像纪录片中出现的数据科学家、数据架构师、数据挖掘师、数据分析师等专家人士，利用大数据技术，破解出那些曾经被认为难以解决的问题，用大数据重新定义世界。

/ 对数据保持敏感将成为未来职场人的分水岭 /

来源 / CPDA数据分析师网　编辑 / 协会会员处 李苗苗　日期 / 2019-09

那么，对数据保持敏感、将数据的价值最大化的数据分析能力就成了职场人提升业务能力或者扩宽职场可能性的必备素质和技能。

拿到任务，一部分人毫无头绪，一脸茫然。一部分人一头扎进数据中，盲目行动！这就会导致...

- 研究问题不清楚
- 大量数据分析方向迷失
- 数据交叉太多无从下手
- 没有清楚的分析思路
- 矛盾的信息
- 关键信息的缺失
- 堆砌太多的数据表格
- 不会善用图表的表达
- 没有清晰逻辑的故事线

好的商业数据分析应该是：

因此，想要掌握数据分析的逻辑、从数据中提炼商业洞察，要掌握系统化的商业分析流程。通过高效表达SCQA模型清楚的框定分析思路。

Situation: 现状分析

首先，根据数据去描述现状，这里的描述要尽量具体；比如：时间、地点、人物、发生了什么现象以及现象达到何种程度。

问题：这张表格可以推断出哪些主要数据发现？

地区	产品	01月		02月		03月		Q1	
		原指标	实际	原指标	实际	原指标	实际	原指标	实际
北区	产品A	152.950	250.520	119.130	156.820	139.900	114.940	411.990	522.280
北区	产品B	11.410	31.260	8980	9570	10.520	13.200	30.910	54.030
北区	产品C	142.420	183.860	111.820	132.230	131.240	96.630	285.480	412.730
北区	产品D	3600	8580	2790	4890	3290	5610	9690	19.080
北区	产品E	1410	1110	1120	1140	1310	990	3840	3240
北区	产品F	23.280	34.410	18.040	21.170	21.250	15.340	62.580	70.920
北区	产品G	31.820	39.070	26.080	32.560	30.170	27.060	88.070	98.680
北区	产品H	290	1890	280	900	300	400	880	3190
北区	产品I	3020	2280	2600	-	30.700		8690	2280
北区合计		370.210	552.980	290.850	359.290	341.060	274.160	1002.120	1186.440
南区	产品A	110.200	199.230	94.830	82.850	104.830	54.640	309.860	346.730
南区	产品B	8180	4120	7470	10.390	8820	9290	24.470	23.810
南区	产品C	96.870	175.330	82.950	67.840	92.170	57.750	271.990	300.930
南区	产品D	3310	7210	2790	4210	3110	4670	9220	16.090
南区	产品E	1230	200	1070	70	1180	950	3480	1220
南区	产品F	30.700	63.880	26.120	35.590	29.500	26.350	86.320	125.820
南区	产品G	21.540	38.080	19.610	18.430	21.380	19.510	62.520	76.020
南区	产品H	11.020	19.400	9570	5590	10.570	7900	31.160	32.890
南区	产品I	2710	-	2600	-	3380		8700	
南区合计		285.750	507.460	247.010	224.980	274.960	191.070	807.720	923.510
全国		1524.330	2443.000	1216.190	1375.920	1359.230	1029.980	4099.750	4848.910

答案

全国Q1的销售表现优异：超出指标18%；

北区超出指标18%，与全国水平持平；

南区表现较弱，超出指标14%，与全国水平比较有差异；

南区未达到全国平均水平是由于2月和3月未达标，主要问题是果汁和瓶装水的销售表现欠佳；

我们要高度注意：Q1全国的销量在下滑；

但是，数据中发现的现状，并不能成为您的商业分析问题。现状数据只是确认发生的事实，我们必须要问“那又怎么样”？这才是将现状直接连接到“冲突”和“问题”，从而寻找答案。

怎样筛选出应该进一步分析的问题？众多的现状数据点是否有值得进一步分析的商业问题？判断依据是什么？这就需要从以下三个维度分析：数据影响力、商业意义和报告受众。

Complication: 探索冲突

冲突是问题或者障碍，是现状与期望的落差。这个落差并非只是不良状态，也有可能是因为超过了预期，证明这是个千载难逢的机会。但冲突的寻找一定要对照“标杆”：

简单来说，例如A公司历年的销售成长为25%, 5年内翻了3 翻（这就是现状）；但是品类与竞争对手同期增长更快，达到了60%（这才是冲突）。

所以如果你是销售经理就需要分析：我们的业务出了什么问题？怎么会这么糟糕？

Question: 定义问题

针对冲突提出疑问，这是一个将冲突转变为解决问题的过程。

S（现状）	C（冲突）	Q（问题）
过去30年来一直采用屡试不爽的方法销售产品，每年增长10%	季度预测显示销售额降低10%，而非增长，预计年底无法达成目标	点击下方空白处查看答案
市场上整体品类年增长30%	我们作为这个品类中的品牌之增长15%	点击下方空白处查看答案
竞争对手10个单品取得30%的市场份额	我们用30个单品才取得15%的市场份额	点击下方空白处查看答案

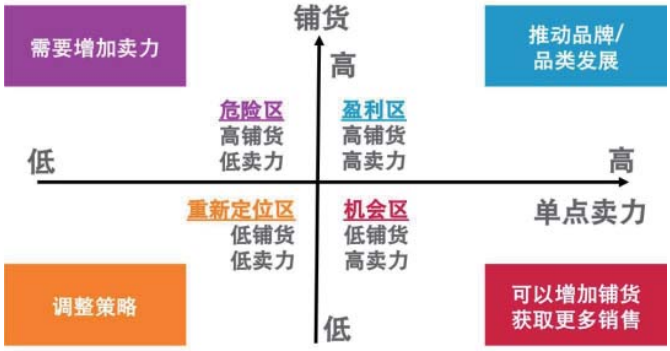
4W思维：When (时间推进法)、Where (地点对比法）、What （漏斗分析法）、Why （因果结构树）是定义问题较为常用的分析逻辑。

以WHERE（地点对比法）为例，你需要关注随地点的对比来解释变化：

以WHAT（漏斗分析法）为例，你需要关注谁在推动品类表现：

以WHY（因果结构树）为例，你需要关注销售恒等式：

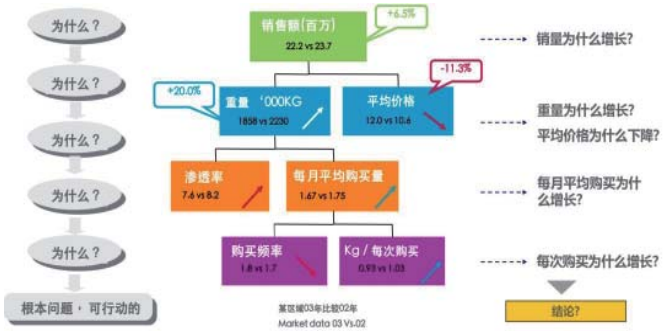
以及“铺货与卖力”矩阵图：



Answer: 解决方案

透析数据将结论转换为洞察和行动方案，强调机会并产生行动才是最终目的！这其中可以通过5个为什么去挖掘表面问题背后的根本原因，4个指标来评估洞察力是否有价值。

5个为什么分析



4个指标评估

获取大量的数据，从数据沙砾中找寻到“富矿”远比数据本身更具意义，透过数据获取更加全面、精准、实时的商业洞察和决策指导才是关键。

/ 稳准狠！大数据助海关精准打击“洋垃圾”走私 /

作者 / CPDA数据分析师 北京 金敏　编辑 / 协会会员处 李苗苗　日期 / 2019-09

知道什么是“洋垃圾”吗?

先来普及一下相关知识!

洋垃圾：指以走私、夹带等方式进口国家禁止进口的固体废物或未经许可擅自进口属于限制进口的固体废物。

这些“洋垃圾”进口到中国后，往往夹杂着很多有毒有害物质，在处理过程中会导致地方河水、土地、空气被污染。

2018年，党中央和国务院提出，全面禁止“洋垃圾”入境，严厉打击走私，大幅减少固体废物进口种类和数量，力争2020年年底前基本实现固体废物零进口。

为落实党中央和国务院的政策，维护国家生态环境安全，中国海关作为阻拦“洋垃圾”流入国内的第一道关键屏障，迅速采取措施，组织开展了打击“洋垃圾”走私的一系列行动，并取得了显著成效。在运用传统手段保持打击走私高压态势以外，中国海关还亮出了一张创新的“科技牌”，通过人工智能技术，让计算机实现对“洋垃圾”的自动甄别和处置，将“洋垃圾”更为有效的拦截在国门之外，切实维护国门安全。

这项创新工作由海关总署风险司、科技司牵头，信息中

心、数据中心及全国各关120名业务、技术骨干参与，联合业界知名机构，开展了一场创新度高、难度性大、专业性强的大数据应用“百日攻关”，通过构建“洋垃圾伪瞒报风险甄别和处置模型”，切实提高海关现场口岸监管能力。

该模型以海关内外部数据为基础，从报关单、物流、商品、企业等多个方面进行交叉比对分析，挖掘行为特征进行机器学习训练，同时设定了多个专家经验风险规则。

洋垃圾伪瞒报风险甄别和处置模型攻关人员

考虑到该模型本质上是一种反欺诈模型，目标是对伪瞒报行为的微弱特征进行甄别和放大，因此攻关团队在比较了CATBoost、XGBoost、随机森林、GBDT等多种机器学习算法后，最终选择了反欺诈效果较好的算法，大幅提高了模型的预测精度。

2019年3月，“洋垃圾伪瞒报风险甄别和处置模型”正式部署至海关作业系统，在天津、南京、宁波、厦门、青岛、广州、黄埔、深圳、湛江、江门等10个海关现场投入应用，命中报关单上百票，经海关技术中心鉴定，查获多票属于我国禁止进口的固体废物。

打击“洋垃圾”走私是打赢蓝天保卫战、建设美丽中国的重要一环，信息中心将继续按照总署统一部署，运用人工智能、大数据等技术手段，为精准打击“洋垃圾”走私、保障国家生态环境安全和人民群众的生命健康安全提供技术支持，坚决对“洋垃圾”说“不”！。

/ 基于移动云教学平台的学情数据分析实证研究 /

作者 / CPDA数据分析师 广东 刘云鹏　编辑 / 协会会员处 李苗苗　日期 / 2019-09

在“移动互联网+”时代来临之际，切实开展真正的翻转课堂教学，将会开启大数据时代高等教育教学改革的新天地，每个教师都可以是自己课堂的大数据生产者。本研究通过移动学习环境下，开展云班课进行《动态网站设计》的混合式教学实践，并通过云班课收集的详细数据，利用“Datahoop”大数据分析平台，对学生的学习状况进行分类，分析学生的学习动机，探索基于网络的学生学习心理，为教师开展网络教学、制作满足学生需求的课程资源和学生考核管理办法等提供数据和理论支撑。

一、前言

“大智移云”时代的高等教育，提高教育质量必然是重中之重。保证课堂教学质量、提升学生的综合素质更是教育信息化发展的核心目标。而教育领域的大数据研究，更应该追本溯源，努力在学科本位、知识本位、课程本位上利用大数据技术与方法，深度解析学生的课堂表现与学习效果，并分析学生的学习动机，从而产生积极的教育影响，帮助教师挖掘教学过程中潜在的影响因素，并加以针对性的改善。

二、研究方法

（一）教育大数据在学情方面的相关研究

2012年美国智库布鲁金斯研究院Darrell M. West最早提出，在教育领域中，可通过收集学生使用、互动信息，课程信息获取相关的教学数据如：分类（Systematic）数据、实时(Real-Time)数据、利用数据监管(Data Curation），高校行政部门可以获知相关教育管理信息，从而进行预测评估学校的各类教学信息，并通过数据可视化的方法，直观的显示给教育决策者以及教师，以便取得进一步的教育决策，最大化的监控学情。同年，美国教育部发布了《通过教育数据挖掘和学习分析促进教与学》（Enhancing Teaching and Learning through Educational Data Mining and Learning Analytics）提出了

教育领域数据挖掘的四个范式目标，即：（1）开发学习者范式，通过对学习者的知识体系结构、学习动机、元认知、学习态度来构建范式，并预测学习者的学习状况及未来成绩。（2）开发相关课堂教学范式，最大化优化教学内容和教学方法。（3）开发各类教学软件运用范式，方便教育数据的采集。（4）综合考量学习者、课堂教学、软件运用等因素下，大数据时代的有效学习范式。

在以上范式目标下，教育大数据的研究主要可以采用以下几种方法Baker（2011）提出了：（1）趋势分析（Prediction）。通过历史数据进行多个变量的预测模式，如研究者通过在线学习环境中学习者参与在线讨论的情况 、 测试情况等，预测学习者在该门课程的学习中是否有失败的风险。（2）聚类分析（Clustering）。根据数据特性，将一个完整的数据集分成不同的子集 ，例如，研究者根据学习者在在线学习环境中学习困难 、交互模式等将学习者分成不同的群组，进而为不同的群组提供合适的学习资源和组织合适的学习活动。（3）关系挖掘(Relationship mining)。 探索数据集中各变量之间的相关关系，并将相关关系作为一条规则进行编码，例如，研究者利用关系挖掘，探索在线学习环境中学习者学习活动和学习成绩的相关关系，进而用于改进学习内容呈现方式和序列，以及在线教学方法。（4）自然语言转化（Distillation for human judgment）。用一种便于人类理解的方式描述数据，以便人们能够快速地判断和区分数据特征，该方法主要以可视化数据分析技术为主，用以改善机器学习模型。（5）模型构建（Discovery with models）。 通过对数据集的聚类 、相关关系挖掘等过程 ，构建供未来分析的有效现象解释模型。

Cristobal（2013）提出自适应学习和个性化学习将会成为一种新的学习范式，而数据对这种新范式的实践起着至关重要的作用。对于个性化教育，人们需要确定学生的相关数据，了解学习提升的真正要素，从而提供有针对性的教学。教育数据的挖掘，就是开发、研究和应用计算机方法从教育大数据中发现学习模式和特征,并且提出了数据挖掘的典型步骤包括：数据采集、数据预处理（如，数据“清理”）、数据挖掘和结果验证。机器智能学习系统可以保存人机交互所产生的详细日志，包括按键点击、眼动跟踪和视频数据。教育数据的挖掘即通过开发、研究并应用计算机智能化的方法，来检测教育大数据的模式。这些数据不仅包括学生个体与智能系统之间的交互数据（如，导向行为、互动练习等），也包括来自于学生之间的协作（如，文本聊天）、管理数据（如，学校管理、教师管理等）以及学生个体情况数据（如，性别、年龄、学校成绩、学习轨迹等）。学生的情感数据（如，动机、情绪状态等）是教育数据挖掘的重点，这些数据可以从生理传感（例如，面部表情、坐姿等）中推断出来。

基于以上的研究，本研究采用的方法主要是通过《动态网站设计》课堂教学实践中，依靠蓝墨云班课软件积累的自然教学数据，通过进行趋势分析、聚类分析等方法，进行学情分析。最后再通过数据可视化，解读学情背后的学生学习动机、学习心理。

（二）数据来源

结合教学实际，该数据选取的研究对象是16级的两个不同专业的本科班，每班人数约在30人左右，并通过同一学期的混合式翻转课堂教学实践。研究对象的课程成绩是由期末成绩+平时成绩两部分构成。总成绩=期末成绩×70%＋平时成绩×30%。本研究的数据来源于蓝墨云班课教学的动态网站设计课程的两个教学班A班与B班，通过数据对比进行研究。

三、成绩趋势分析

参考国内外研究，结合独立学院“动态网站设计”的教学实际，课程团队构建了“动态网站设计”形成性评价方案。将“动态网站设计”学期成绩分为“平时考核成绩“与”期末考核成绩”两部分。比例为30%与70&。前者包含了六项活动分别为：非视频资源学习10%；签到10%；测试15%；讨论答疑5%；头脑风暴5%；投票问卷10%；作业/小组任务15% 课堂表现15%；被老师点赞加分5%。

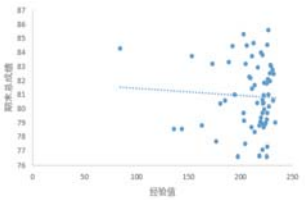


图 1　A 班经验值与总成绩散点图
Figure 1 Scatter Diagram of Experience Value and Total Achievement of Class A

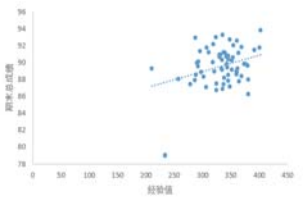


图 2　B 班经验值与总成绩散点图
Figure 2 Scatter Diagram of Experience Value and Total Achievement of Class B

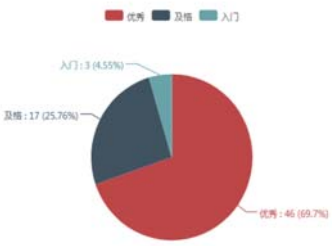


图 3 A 班经验值分布
Figure 3 Distribution of Experience Value for Class A

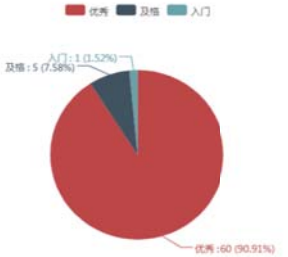


图 4 B 班经验值分布
Figure 4 Distribution of Experience Value for Class B

通过期末成绩分布表，我们可以看出A班的学生成绩与B班的学生成绩略有不同，A班成绩较为发散，B班成绩较为集中，为更进一步的研究A、B班的成绩分布，绘制散点图如图1 图2 通过绘制经验值为X轴，总成绩为Y轴的散点图可以清楚的看到两个班的成绩分布与经验值分布之间的关系状况，从散点分布来看，A班的经验值多集中在200-250之间，B班的集中

在250-400之间，再从总成绩来看，A班经验值高的同学期末并不一定高，甚至呈现出经验值越高，总成绩越低的趋势，而B班则呈现出截然不同的趋势，经验值越高，期末总成绩也越高，并且成绩与经验值的分布程度更加集中。为研究这两种分布现象背后是什么原因造成的，有必要研究对两个班进行更进一步的研究。

Table 1 Class a and class b performance distribution table

表 1 A、B 班成绩分布表

分数	A 班人数	B 班人数
100-90 分	35	1
89-80 分	29	40
79-70 分	3	26
69-60 分	1	5
59 分及其以下	4	7

四、学情分析

为挖掘两个班的学情，利用蓝墨云班课导出的学情数据，需要选取有代表性的多个指标，通过聚类，尝试找到影响学生学习成绩的潜在因素。数据采集的经验值和期末成绩通过简单的回归分析，可以得到简单回归分析的两个教学班的指标如表2 所示：

Table 2 A, class b results distribution statistics

表 2 A、B 班成绩分布统计表

项目	A 班	B 班
回归方程	y=0.0194x+83.16	y=-0.0051x + 81.955
R2	0.10	0.00

依据上表可以看出，建立线性回归方程无统计学意义。需要引入新的方法，也就是聚类法，本研究利用当下流行的大数据SAAS平台“DataHoop”来进行聚类测算，该平台的优势是内置丰富的数据分析和数据挖掘算法，能实现算法参数的自动调优和升级，同时包含了适合中国国情的行业应用模型。首先利用蓝墨云班课导出的学情数据可以看出课堂经验值的构成主要如下表所示。

Table 3 A, b class experience value itemized list

表 3 A、B 班经验值分项列表

项目名称	非视频资源学习	签到	测试	讨论答疑	头脑风暴	投票问卷	作业/小组任务	课堂表现	被老师点赞
比例	10%	10%	15%	5%	5%	10%	15%	15%	5%

Table 4 A, class b phase empirical values are divided into correlation matrix

表 4 A、B 班相经验值分相关系阵

	头脑风暴	讨论答疑	同学点赞总经验值	被老师点赞经验值	课堂表现	非视频资源学习	课堂测验得分
头脑风暴	1	0.8694	0.5537	0.9362	-0.0883	0.3177	0.6315
讨论答疑	0.8694	1	0.7293	0.9185	-0.1419	0.42	0.4523
同学点赞总经验值	0.5537	0.7293	1	0.8111	-0.042	0.6865	0.3569
被老师点赞经验值	0.9362	0.9185	0.8111	1	-0.0798	0.513	0.5942
课堂表现	-0.0883	-0.1419	-0.042	-0.0798	1	0.4565	0.0294
非视频资源学习	0.3177	0.42	0.6865	0.513	0.4565	1	0.3483
课堂测验分	0.6315	0.4523	0.3569	0.5942	0.0294	0.3483	1

将A、B班的以上各指标数据直接上传至“DataHoop”，为检验各类分项后所隐含的含义，首先检查变量间的多重共线性，从而能避免结果错误。根据表4可以看出，通过测算相关矩阵，《动态网站设计》该门课程的课堂活动因素存在较高的多重共线性，通过该检验可以发现变量 “头脑风暴”与变量“被老师点赞总经验值”，“点赞总经验值”与变量“被老师点赞总经验值”等变量间存在较高的多重共线性，需要减少变量，并需要为找出下一步影响学业成绩的潜在变量之前进行因子分析，也就是对该数据集进行“降维”见图5

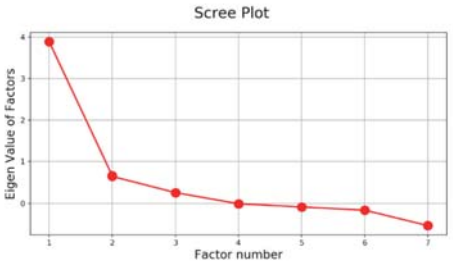


图 5 经验值指标因子碎石图

Figure 5 Gravel Diagram of Empirical Index Factor

Table 5 Contribution Rate of New Variables Generated after Factor Transposition Matrix

表 5 因子转置矩阵后产生的新的变量的贡献率

	F_1	F_2	F_3	F_4
贡献率	0. 5769	0. 2131	0. 5595	0. 1335
累计贡献率	0. 5769	0. 79	0. 5595	0. 693

依据DataHoop检验报告（DataHoop可以在分析时默认进行标准化处理）表5可以看出，A班数据通过因子分析转置后产生的新的变量F1、F2贡献率分别为0.57与0.21 B班数据产生的新的变量F3、F4分别为0.56与0.13. 并对数据进行聚类分析，根据聚类结果分析每一类客户在现有变量上的特征，这里选取平均值作为参考依据。得到聚类分析描述结果为：

Table 6 correlation coefficient arrays (average profile coefficient is 0.67)

表 6 A、B 班相经验值分相关系阵 (平均轮廓系数为 0.67)

类别	1	2	3
样本个数	24	108	19
头脑风暴	-1.25	0.25	0.15
讨论答疑	-0.8	0.41	-1.34
同学点赞总经验值	0.42	0.25	-1.97
被老师点赞经验值	-0.35	-0.1	-0.08
课堂表现	-1.54	0.35	1.03
非视频资源学习	-1.71	0.32	0.32
课堂测验得分	-0.98	0.37	-0.9
讨论答疑	-0.4	0.15	-0.37

依赖于蓝墨云班课软件的数据，在当前成绩下，将两个班151名学生作为研究对象，我们在DataHoop上进行了

K-means聚类分析，依据检验报告的结果， 可将学生依据课堂活动的指标，分为三类。研究结果如表6所示。

第一类学生，样本数24人，在课堂活动中，各项课堂活动得分均较其他两类学生更低，这部分学生课堂活动参与不积极，但 “同学点赞经验值”这项指标最高。 也可以认为在翻转课堂的教学场景下，开展的学生互评存在主观性，学生互评阶段的分数存在打“同情分”，打“感情分因素”的情况，由此可知，此项指标“点赞经验值”在该课程的学生互评过程中并不客观，需要教师设计合理的学生评价指标和统一规范的打分机制。

第二类学生，样本数108人，作为人数最多的类别，各项活动得分比较高，意味着翻转课堂的教学情境下大多数同学在实际教学中都能积极、按时的完成各项课堂活动，并达到考核标准。但由于人为主观因素，也可能造成经验值与成绩稍有偏差，但总体偏差不大。

第三类学生，样本数为19人，通过研究，可以发现该类学生在“讨论答疑”、和“同学点赞的经验值”这几项的得分都很低，但是在“课堂表现”、“非视频资源学习”与“被老师点赞的经验值”指标处，得分都很高。通过个案研究与访谈发现，该类同学属于“学习成绩优秀”的学生，在翻转课堂的学习中，应当充分发挥课堂中“榜样”的力量，并在课堂活动中增加 “答疑解惑”指标项的得分权重，鼓励更多同学，能积极主动的在学习过程中相互帮助，从而形成借助翻转课堂与移动云教学平台创建、积累、完善和分享知识的全新模式。

通过以上研究我们可以发现：在翻转课堂中，利用“蓝墨云班课”获得经验值与学习成绩并不直接相关，是由于经验值的获取来源较多，并且其中某些指标得分较为主观,不能直接用作平时成绩，或是经过加权处理后使用。另外翻转课堂中的教学评价，虽然可以参考“蓝墨云班课”软件设定的各种课堂活动作为评价指标，但其中的指标应当区分为“评价性指标”、“活动性指标” 以便在成绩评定时加以区分，使评价更加客观。

五、 结语

大数据带给我们的是颠覆性观念转变：是所有数据，而不是随机采样；是主体方向，而不是个别精确；是关联关系，而不是因果关系。学情的分析，除了数据的分析及支撑外，还需要我们立体的，多角度的对学情分析结果的使用， 这样就能更有针对性的对学生进行个性化的教学。 我们的教育要让每一个学生都得到成长，以及感受学习的快乐。

基于蓝墨云班课 (Mosoteach) APP的深度学习, 以云平台应用优势作为技术支撑, 充分利用多种软件功能, 使教师、学习者、学习伙伴之间形成交互式的复杂关系, 并能为深度学习者提供多样化的丰富教学资源开展进阶式学习的多种途径, 使深度学习贯穿于课堂上下, 并为教师评价学习者、学习者自我评价或互相评价建立了通道。伴随整个深度学习过程, 产生一系列有益于促进深度学习的直观量化的追踪式评价数据。通过这一学习平台展开的设计谨严又能随机变化的深度学习过程, 能充分激发学习者的内在主动性, 使之以反思、质疑、批判的理性态度对新旧知识不断分析、整合, 促进学习者的表达、沟通, 不断进行新知识体系的建构和更新, 使深度学习得以持久延续。同时, 与真实世界相联系, 将所获理论知识迁移运用于解决实际生活问题。

/ 用Python语言做数据分析基本思路和流程 /

作者 / CPDA数据分析师 湖南 熊恩成　编辑 / 协会会员处 李苗苗　日期 / 2019-09

当下用PYTHON做数据分析实在是太火了！大多数招聘信息里都要求应聘者会使用PYTHON做数据分析。PYTHON语言功能确实很强大，俗称“胶水语言”。那么我们大多数职场人士真的有机会使用到PYTHON吗？或者说会PYTHON语言，是必备技能，还是加分项？

一、客观条件还有不足

熊大在之前分享的《企业数字化转型的关键是什么》里指出，经抽样调查，目前国内大多数企业基本的数字化都还未实现，更不用说业务的数据化和在线化。真实现状是，许多企业使用的依然是一个个独立的系统，每个系统的数据基本上也是“独立的”，没有很好的连接起来，其以物理的状态存储在数据库中，没有发挥出数据应有的价值。

而目前，许多企业已意识到了数据的价值，于是纷纷招聘数据分析方面的人才。然而招聘进来的人发现，日常做数据分析，需要一个系统一个系统地读取数据。仅在各个系统读取

数据可能就会花费大量的时间，后面还要用EXCEL等工具把数据分别预处理好，极大耗费时间成本。

企业招聘进来的数据分析师，本应发挥应有的价值，但实际操作中，其他岗位人员认为读取数据、预处理等工作都应是数据分析人员的，因此，数据分析人员需占用多数时间应付一些基础性、重复性的工作，而真正有价值的事情反而投入不足，无法有效顾及。

并且，实际中令数据分析人员更加痛苦的是，他们会和企业老板提出改进意见，比如把企业的业务数据化、在线化，或者是先把各个子系统数据进行集中存储，建立企业自己的数据仓库，开发自动化数据产品，提高效率等方案。但是，这些都需要花费大量时间和资金，短期内还不一定有成效。企业老板会考虑，到底值不值得投入？很多时候，可能自己也没有思路，企业里也没相应的人才，所以干脆就拖着，到时再看……

二、是必备技能，还是加分项

鉴于当下这种现状，对于想从事数据分析的朋友来说，要么力争去头部企业，因为他们的数据丰富，PYTHON的用武之地更大，因此这就是必备技能。而对于无法进入大厂的朋友来说，会PYTHON可能仅仅只是个加分项，因为使用的机会并不多。当然我们要用发展的眼光来看待，企业数字化转型也是一个逐步实现的过程，不可能一蹴而就。未来，PYTHON的使用机会很多，因此学会PYTHON还是很有必要的，前提是在你条件允许的情况下去学习。

对于那些已工作多年，现在要从事数据分析方面的业务的朋友，建议你把数据分析的工具熟练使用，比如EXCEL、SPSS等。由于有工作经验，则需要建立自己的数据分析思维，掌握必要的数据分析方法。熊大自己的经验是，自从系统学习了数据分析方面的思维和方法，再结合自身的工作经验，现在已能很好地把数据分析技能应用在实际业务场景中，虽然许多东西还在不断完善和改进中。

三、基本思路和流程

对于那些做过实际数据分析业务的朋友，应该对数据分析的基本思路和流程都比较熟悉，也初步掌握了数据分析的思维和框架。比如能够熟练使用SPSS、SAS、EXCEL等分析工具完成中小型的数据分析项目。如果是用Python语言，该怎么做数据分析呢?

基本思路和流程如下:

1、业务目标的确定。这点与分析工具和语言没有直接关系，这里不做过多阐述。核心是你想解决什么实际问题？必须要先明确。

2、数据的获取。我们平时用SPSS和EXCEL做数据分析的时候，都需要事先把数据从系统里导出，然后再导入到分析工具里，再进行下一步。相当于是把要分析的数据“装入”到分析工具里进行加工处理。

若用PYTHON，如何操作呢？原理基本一样，直接在程序运行环境里写一行程序代码即可。当年做PHP开发，程序文件里也是开头用几行PHP代码实现了网站应用与数据的连接。理论上这相当于省去了数据的导出导入步骤，节省了时间。

个人认为掌握这个思路非常重要。当你真正面临“大数据”时，把数据用传统的方法“导入导出”是不太现实的。你需要把自己置于一个虚拟的空间里（云上），想象下，你可操作的“大数据”在云空间的A处，“大数据”的分析框架在B处，要做分析，就需要把A和B使用PYTHON程序进行连接，把数据“装入到”大数据分析框架中，再借助云的高强计算能力，实现对“大数据”的快速实时分析。看，原理是不是差不多?

3、数据的查看。我们平时用EXCEL打开数据表时，可以直观的查看里面的数据情况，比如字段有哪些？有多少数

据量？数据的质量怎么样？缺失值或空值多不多？如果用PYTHON查看的话也很简单，写几行代码，运行一下就可以知道表中数据的所有情况。再比如对“异常值”的检测，使用PYTHON，同样可以用“箱线图”显示出来。

个人体会，当数据量不大的时候，使用工具和程序没什么差别。然而一旦数据量达到一定量级，我们都知道像EXCEL这类的工具，数据量稍微大点，打开就很慢，甚至会崩溃，这时使用PYHON程序的优势就显现出来了。

4、数据的预处理。这点也很好理解，例如我们用EXCEL做数据的拆分、合并或者转换，用PYTHON是同样的原理，比如用“split()”做数据拆分，只需要写一行代码即可对字段所对应的数据进行拆分。其他都是类似的操作。是不是很简单？

我们都知道数据分析工具也无非是某些程序语言编写的，无论是EXCEL、SPSS还是SAS。只不过这些工具是通过界面的操作，把“人机对话”的难度大大降低。拿智能手机来说，以后的手机只会更加简单和实用，复杂的机型肯定会被抛弃。

最后总结下，最近几年内对于从事大数据的人来说，使用好EXCEL、SPSS就能应对日常大部分数据分析工作。但是当企业数据量在快速增多的情况下，就需要学会使用更先进的工具来实现。大数据分析师，需要熟悉云数据库、云存储、云计算及各类算法的使用场景，还要掌握大数据分析框架，使自己真正达到能“指挥千军万马”的境界。

对于做数据分析的朋友来说，我们也只需要掌握PYTHON语言中与数据分析紧密相关的几个包，不必要全部都学。PYTHON虽然入门比较简单，但真正用好它还需要花费一定时间。因此我们当务之急要做的，常用的工具会使用，PYTHON常用的函数包尽量会用，工作中遇到问题，哪个能快速解决就使用哪个，最终目的是解决问题。

熊大认为使用语言与操作工具做数据分析的基本思路和流程是差不多的，无论使用任何“武器”，只要“能消灭敌人就好”，前提是我们对每一种武器都要熟练使用，并且能高效地解决问题。

/ 运用数据分析方法提升电信存量用户捆绑率的研究 /

作者 / CPDA数据分析师 武汉 李毅、王月皎、李纯、郭轶、刘莎、王荆　编辑 / 协会会员处 李苗苗　日期 / 2019-09

一、研究现状和发展趋势

存量用户价值问题已越来越受到电信运营商的重视。随着移动通信与其相关行业互相融合，利用新技术、新项目以及应用需求满足当前人们的生活需求，人们享受到互联网时代移动通信技术发展带来的优质用户体验。在新的5G时代，客户价值将成为竞争焦点，对存量价值客户的争夺将成为新赛场。

二、大数据时代存量用户价值提升的方式

（一）存量经营的涵义

存量经营囊括了运营商除拓展新客户外的大部分经营活动，除了前端的市场活动，还包括后端的支撑活动（如网建、网优、IT系统完善等）。表现为运营商针对现有客户，以提升客户忠诚度、释放客户价值为目的的一系列经营方针和策略，主要是通过客户信息挖掘、精细化管理、差异化服务来实现客户保有和价值提升，体现出运营商对流失率的控制，除了要留住客户，还要对客户的价值进行挖掘和提升。随着用户市场日趋饱和，用户增量市场的空间不断减少、增速日益放缓，运营商对存量市场的依赖程度将不断增加。

（二）精准营销流程

精准营销不但提高了营销的准确率和命中率，同时也提高了服务水平。基于运营商的用户大数据，通过大数据挖掘、机器学习等方式进行数据价值的获取，具体营销流程见下图。

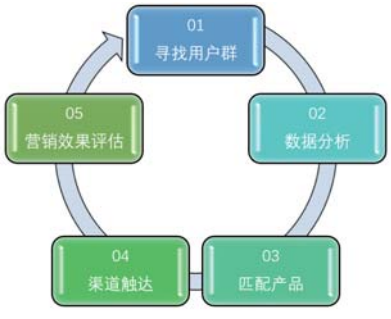


图1：精准营销流程

三、基于聚类分析的客户细分模型建立与评估

细分客户可以帮助运营商更加精准的了解客户的消费行为，了解客户的特征。对客户按照一定的标准进行划分，实现

对市场精准的分割，实现精准营销。

(一)提取用户画像

用户画像即商业目的下用户标签的集合。运营商制定自有的标签体系，并对用户的语音和流量使用情况进行数据统计和分析，从而确定用户所匹配的相应标签。用户标签可以分为两部分，一是自然标签，包括用户的基础信息、性别、年龄、家庭住址等，二是用根据用户的基础信息和行为数据的归纳和分析而来的特征标签，如最常用的APP、最喜欢的电商网站等等。根据用户的基础信息和行为信息，对用户进行360度的属性特征和行为偏好画像。用户标签产品能全方位的了解用户行为特征，为锁定潜在目标用户群、营销决策等提供数据支撑基础。

（二）数据维度与样本量的确定

本文所用实验取公司数据服务支撑平台的用户数据为样本。根据电信用户特点，针对性的选取了用户属性，剔除了无效的属性，因数据维度较低，无需对数据做降维处理。根据电信业务经营的常识，结合本次实验的需要，在数据选取时将遵从以下两条原则：

1.在不损坏数据完整性的情况下，合并意义相近或相同的字段；

2.与本实验无关的字段进行剔除，裁剪掉无效的字段。

本实验将随机抽取100000户样本用户（在网6个月以上），剔除无效的或缺省值较多的字段，有效数据占样本整体量的98.05%，最后得到98050户数据作为样本参与客户细分模型建立。

（三）用户分类

用户中包含两部分，一部分是订阅过流量包的用户，一类是非订阅用户，计算二者比例。

拟符合流量包订购条件样本62950户，另4770户已为冰激凌套餐,10327户近半年无流量需求。

流量包	层级	国内	赠送省内	样本数	备注
半年包	40 半年包	-	3G	14504	
流量月包	10 元	100 M	-	18300	流量 1 元卡(OCS)12 户、ARPU 值小于等于 10 元 8 户
	20 元	300 M	-	13005	
	30 元	500 M	1000M+2G 省内*12 个月促销	9700	
	50 元	1G	5G	4652	
	70 元	2G	10G	2599	腾讯大王卡等 21 资费 35 户，其中 ARPU 值小于 60 元的 30 户。
	100 元	3G	15G	190	

（四）数据建模

根据数据字段及样本的选取规则，从上述的98050个数据中筛选过滤，得到下面部分字段的客户数据信息。有部分字段在系统中直接提取出来是无法使用的，经过转化求和等方式进行了加工，才得到以下内容，主要字段列举如下：

表1：客户画像数据字段

字段名	数据类型	说明
用户识别号	Text	用户唯一识别码
用户号码	Text	用户服务号码
在网时长	Number	用户在网时长
用户产品名称	Text	用户产品名称
是否订购流量包	Text	用户是否订购流量包
使用网络类型	Number	用户使用网络类型
手机上网费	Number	用户 6 个月平均上网费用
ARPU 值	Number	用户 6 个月平均 ARPU 值
国内语音时长	Number	用户 6 个月国内语音平均时长
国内流量	Number	用户 6 个月平均国内流量

剔除存在缺失值、空值样本后，选取有通信行为的客户数据98050条作为聚类目标对象集，部分客户数据如下表所示（本文涉及的数据均经过特殊加工处理）

表2客户使用行为数据表

用户标识	在网时长	产品名称	流量包	使用网络	上网费（元）	ARPU值	国内语音时长	国内流量（MB）
155**588	99	2	1	2G	0	9.65	10	0
130**362	148	4	1	3G	0	140	24	0
131**856	5	4	0	3G	3.75	53.1	57	1.1
130**912	37	3	0	4G	34	120.5	0	1899
156**872	42	1	1	2G	13.6	30.5	138	707
131**949	46	4	0	4G	0	10.5	63	267
155**603	75	2	1	4G	86	100.2	294	4300
186**686	42	1	1	4G	170	230.2	217	8905
155**690	89	3	1	3G	6	138.1	23	6.3
131**698	24	2	0	2G	0	36.8	31	0

（五）客户细分模型建立与评估

使用Datahoop平台，将套餐名称、网龄、ARPU值、国内语音（分钟）、月均国内流量（G）、age、月均消费、免费流量、收费流量、收费和流量费等变量都放在X轴导入系统进行分析。

使用“异常值分析”，有39个异常值占比不高，可不做处理。

使用“相关系数矩阵”工具计算相关性，可知免费流量和收费流量系数大于0.9，属于强相关，其他均小于0.7，因此将免费流量和收费流量的比值（免费流量/收费流量）作为新的变量，与网龄、ARPU值、月均国内语音（分钟）、月均国内流量（G）一起作为变量进行下一步建模。

将上述五个变量再次导入Datahoop平台进行“相关系数矩阵”。结果显示其相关性均小于0.8，不存在强相关，可以使用聚类分析来对这些用户进行分类。

使用聚类中的K-means计算方法，在Datahoop平台中建立数据流，将聚类系数分别设定为3、4、5，来检测何种分类下情况最佳。

结果显示，在聚类系数为4时，平均轮廓系数为0.6537，此时分类效果最佳。

k-value	silhouette coefficient
3	0.4539
4	0.6537
5	0.6497

图1：各种分类下的轮廓系数结果

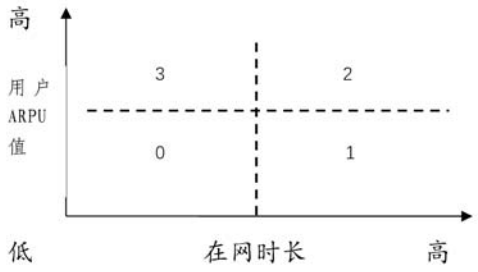


图2 建模过程图

(0为不常使用用户;1为中端日常用户;2为中端商务用户;3为高端商务用户)

类别	ARPU 值(Y)	Domestic Voice(S)	Domestic flows(M)	Network age(M)	客户分类
0	19.62	58	1.13	6	不常使用用户
1	45.72	1823	325.72	15	中端日常用户
2	98.9	6582	4321.69	18	中端商务用户
3	135.5	31257	10732.78	8	高端商务用户

通过聚类模型分析结果如下：

0类用户为不常使用用户，可以尝试推荐10元100M-30元500M流量包，培养用户使用流量习惯，提高用户在网时长，目的在于挽留用户继续使用。

1类用户为中端日常用户，适合推荐存费送电子券产品预存120元赠送240元，12个月。

2类用户为中端商务用户，适合推荐存费送电子券产品预存240元赠送480元，12个月

3类用户为高端商务用户， 适合推荐用户迁转为畅越冰激凌套餐。

通过聚类分析后得到的分类结果，还是比较符合用户实际使用情况的。接下来就运用这种分类方式对我们将要推荐业务的用户数据进行分类并进行实践。

四、存量用户价值提升实证分析及研究

取出一批待维系的用户数据，提取他们的属性，即建模过程张选取的属性，将这些数据导入datahoop平台，使用聚类k-means算法进行分类。按照分类的结果对不同类的用户进行不同产品的推荐。

建模目标在于，确定用户的个人属性和消费能力对其订购相应业务的影响力大小，以便于对其他非订购用户进行业务适配，并进行相应推荐。

由于研究能提供的触点渠道存在局限性，本次采用了电话营销的传统方式进行渠道触达。电话营销方式的客户感知是远低于通过APP或网上营业厅等渠道的客户主动行为转化的感知。据统计，前期存费送电子券、套餐迁转和流量包订购为1.3%，通过大数据精准营销转化率达到了2.3%，提升了个1个PP。从实证结果来看，模型的运行结果实际可行，可以运用于日常工作中。

五、结束语

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，使我们建立了数据分析思维、熟悉了数据分析流程、掌握了数据分析方法和技能，并结合工作需求进行实际应用的研究，但由于资源条件、设备工具和水平所限，本文在数据分析，模型选择及描述等方面尚有很多不足，需要在未来研究中进一步验证价值，提升研究的有效性和科学性及工作实际应用性。通过数据分析方法，在一定程度上拓展相关领域的分析视角和方法，充分挖掘数据价值，产生数据效益，促进行业的发展。

/ 创业项目发展关键，数据的重要性 /

作者 / 陕西诚合广信数据分析师事务所　编辑 / 协会会员处 李苗苗　日期 / 2019-09

诚合广信是一家专业从事深化数据分析研究的事务所，主要为企业提供专业的数据分析和科学的决策数据。事务所坐落于美丽的十三朝古都陕西西安，员工来自五湖四海，秉承着与时俱进、开拓创新的员工精神，坚持诚信为先、以人为本、客户为尊的核心价值观，确立以员工为根本，以管理为基础，以客户为中心，以社会效益为目标，建立高效专业的数据分析团队为发展方向，打造了数据行业一流的服务体系。

事务所设有数据采集中心、项目部、质量控制中心等研究部门以及咨询部、行政部、人力资源部、财务部等日常事务部门，共计三十余人。事务所有一流的分析师数十人，其中很多分析师做过国家级重点项目和大型企业的数据分析。

诚合广信数据分析师事务所成立前后已经为广大客户提供了各种深化的数据分析服务，包括《福建漳平年产8000吨间接氧化锌、18000吨活性氧化锌项目数据分析》、《核桃和油用牡丹及中药深加工产业化建设数据分析》、《芜湖年处理10万吨废旧高分子材料资源循环利用项目数据分析》、《四川新材料医用贴生产项目数据分析》等就客户项目关键问题指出并做出适应企业发展的调整，给出了最佳的解决方案，让客户认识到了数据分析的价值、战略规划和管理意识上的不足，不止给客户提供了一份资料，更是给客户在未来的发展上铺平了道路。

诚合广信数据分析师事务所自成立以来，一直牢固树立“以人为本、客户为尊”的宗旨，坚持合理、有效、客观、公正的数据分析原则，严格遵守国家相关法律法规及数据分析人员职业道德，开展数据分析等相关业务，为客户提供了有效、准确、有价值的数据依据，制作出很多对客户有帮助的数据资料，让客户在发展的道路上顺利前行，得到了广大客户的一致好评。

我事务所诚邀全国数据分析师兄弟单位、全国各地数据分析师朋友，以及对各行各业数据研究有需求的单位，到我事务所交流、考察、洽谈合作。

联系方式：陕西诚合广信数据分析师事务所
联系人：张女士
联系方式：029-62710005 , 13072909507
地址：陕西省西安市碑林区雁塔路中段甲字16号
东新科贸C楼4层425室
网址：www.chgxsj.com

/ 江苏某大型三甲医院数据分析案例 /

作者 / 上海加睿数据分析师事务所　编辑 / 协会会员处 李苗苗　日期 / 2019-09

一、项目背景

医院为某地区三级甲等综合医院，现有省级临床重点专科23个，市级临床重点专科20个。

医院经过多年的发展，信息化水平在全国属于较高水平，近年来为了将信息化建设提高到更高水平，医院决定进行信息化改进，通过制定以电子病历评级标准6级水平为建设目标，实现以评助建的效果，改善医院医疗信息化水平。在此过程中，发现了很多数据质量及数据标准相关问题，主要包括：（1）医院内部各业务系统数据缺乏统一的数据标准和规范，导致部分系统数据交互数据不能识别问题，影响了临床数据应用的效率。（2）各业务系统及数据中心数据质量不高，导致数据无法发出最大应用价值。（3）病人基本信息及患者的临床资料，在医联体中医院内部无法很好的识别及共享，对医联体业务的有效开展，形成了很大的制约。

这些问题严重影响了医院进行基于数据的更深入的应用，以及医联体的建设效果，如何提高数据标准及数据质量，实现医院数据质量的持续改进，是医院一项非常急迫的任务。

二、医院数据分析效果

通过数据分析平台的建设，提高了医院内部信息管理的水平，实现了数据资产化管理；并通过PDCA的数据质量管理理念，实现了数据质量的持续改进，为临床业务的改善及运营的精细化管理带来了坚实的基础

1、实现资产化管理，提高信息管理水平

（1）满足国家六级电子病历评审

对支持业务需求的数据进行全面质量管控，通过数据质量相关管理办法，参照《电子病历系统应用水平分级评价方法及标准（2018版）》，满足该医院六级国家电子病历评审的数据质量管理要求。通过组织流程、评价考核规则的制定，及时发现并解决数据质量问题，提升数据的完整性、及时性、准确性及一致性，提升医院业务价值。

（2）实现医院数据资产的盘点

通过整合、维护数据分析平台，记录医院各业务系统数据库资产的表、视图、索引、存储、报表、指标等数据资产信息。实现医院全数据的可视化管理，形成医院数据资产。

通过定时任务检查系统，发现资产（表结构、视图、存储…）的变化，进行预警，杜绝不合规的操作情况。实现医院数据资产的合理管控、做到医院数据资产的使用透明。

通过定期给医院数据资产进行评估、体检，对系统升级前后资产对比，形成版本管理体系，方便医院回退和版本功能对比。全面实现数据资产的可管、可控、可用、可信。

（3）构建医院数据地图

通过维护数据库表、字段的关联关系和数据追溯的血缘关系，以业务层、缓存层、数据仓库层、数据集市层追朔关系，形成数据地图模型。帮助医院快速定位数据问题，分析数据流向，明确数据的来源、去向。系统提供完整细致的血缘分

析，对问题的节点进行回溯，分析其处理路径上可能存在的问题以及相关影响范围，对医院的数据变更提供保障。

2、对临床业务的持续改进

（1）实现各系统患者数据的一致性

通过数据质量的完整性规则校验，对门诊就诊、住院登记、诊断录入、医嘱录入、护理记录、检查、检验、手术等患者基本信息（如患者标识，姓名，地址，联系方式等）的一致性进行稽查。在稽查过程中下发现诸多问题，院方当即下发整改方案、通过系统改造及加强管理方式督促各科室医生进行改善，提高医院数据质量，提升医院就诊水平。

（2）手术及诊断的一致性完善

在数据质量的核查过程中，发现在医生、护士、手麻等系统中部分患者诊断及手术名称不一致情况，给医护人员带来了不小的困惑，也给病人的后续治疗动作带来了不小的安全隐患；经过分析，这个问题的因为相应的系统中录入诊断和手术的界面，支持手工输入，且系统之间相应数据做了变化，无法同步到其他系统中； 医院根据实际情况加强了对医护人员的系统配置，并召集了各系统厂商，进行了通过完善相应的信息系统，保障了各个系统中的数据的一致性。

（3）纠正了不规范的医疗行为

医院相关业务存在一定的连续性及关联性，示例如下：

医生站手术申请单数 <= 手术记录数 <= 麻醉记录数

医生站检查申请单数 <= 检查预约登记数 <= 检查报告记录数

医生站检验申请单数 <= 检验预约登记数 <= 检验报告记录数

医生站治疗申请单数 <= 治疗记录数 <= 治疗报告数

从数据的逻辑性上来说，不同业务系统中的相关数据，下游业务系统的相关数据应该与上游数据可对照。例如，医生站检验申请单数与检验报告记录数，有申请单不一定有报告，但是有报告的理论上应都要有申请单。

通过数据分析平台关联业务数据的检查，发现部分检查、检验报告无申请单关联，通过分析，找出问题原因是：部分患者或者是院内职工在没有申请直接做检查、检验项目，院方根据此检查质量报告及时下发整改方案，明确医务人员按流程按规章处理就医事宜，纠正了这些不规范的医疗行为，减少医院的经济损失。

三、医联体数据分析效果

1、实现患者的统一识别

同一患者，在医联体内部的识别码是不同的，在各医院上传的各类临床资料中只是保存了各自生成的患者ID，导致不同医院就诊的患者资料无法全部识别，通过建立医联体内患者主索引，实现医联体内部的患者识别，并通过数据质量管理系统，实现了对各医院上传的数据的完整性，有效性等稽查，通过对各医疗机构的上传程序的改造和完善，保证了患者主索引能够得到很好的实现，为患者的双向转诊，转检等业务的开展提供了很好的数据基础。

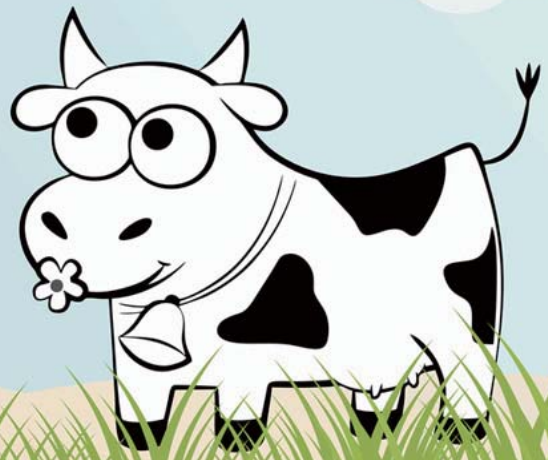
2、实现双向转诊及患者数据的共享

在建立医联体建设过程中，各个医院的临床资料虽然能够上传到医联体数据中心，但由于各个医疗机构内部的数据标准不一致，医院的数据在其他医疗机构内部无法得到很好的识别，导致患者在医院的就诊数据，在其他院内无法被很好应用，无法保证患者的临床资料的连续性，通过医联体数据分析平台的建设，建立了一套医联体内部数据标准，各医疗机构按照标准进行数据转换及数据交互，并通过数据质量系统实现数据的质量的稽查，保障数据的标准及质量。当患者在一个医院挂号时，相应的数据将会被采集到医院内部数据中心，通过患者全息视图实现患者临床资料的整合，医生可以查看这个病人在其他医疗机构内部的就诊资料，提高医生的就诊效率。

联系方式：上海加睿数据分析师事务所
联系人：叶女士
联系电话: 13917326043（同微信号）
地址：上海市闵行区七宝镇新龙路1399号宝龙城T4座1003室

/ 奶牛和企业舆情数据分析 /

作者 / 山东智谷数据分析师事务所 李春燕 王烁 苏航 编辑 / 协会会员处 李苗苗 日期 / 2019-09



随着数据时代的到来，越来越多的传统行业已经意识到了数据分析的重要性，例如奶牛养殖企业和公安部门，他们将行业数据与Python或R语言数据分析结合，充分发挥了数据价值，提升了业务水平。

案例一：奶牛遗传育种

1 牛奶检测结果数据分析

相关科研机构从每头奶牛产生的牛奶中抽取样品形成奶样，奶样需经专业的仪器检测12个指标，这些指标对于奶牛场来说只是一堆数字，我们要做的就是让数字会说话，透过数字展示奶牛场的养殖优缺点，为奶牛的科学养殖提供依据。

首先，观察数据量和数据格式，选择分析工具。获取到的数据是Excel表格形式，数据的格式都是数值型且不存在异常值，数据量不大（在100-1000条之间），可以直接用Python处理。数据样本见图1。

牛号	乳脂率	蛋白率	乳糖率	干物质	Cond.	H.Index	乳尿素	体细胞	ZValue	PhaMean
1569	3.33	2.73	4.91	11.13	1005.9	0.76	21.9	141	2.55	70
1440	1.25	3.28	5.08	9.53	939.1	0.65	28.1	154	2.29	69
1401	1.4	3.38	5.11	9.82	957.3	0.65	24.2	605	2.88	65
1720	2.34	2.74	4.99	10.11	1020.5	0.72	25.7	17	0	0
1723	3.18	2.89	5.14	11.23	884.4	0.68	22	74	2.64	66
14006	1.66	3.26	4.89	9.82	1032.1	0.72	28.9	35	0	0
1701	2.86	3.04	4.95	10.92	936	0.74	26.3	235	2.49	70
1463	1.61	2.91	5.36	9.82	907.9	0.62	23.6	53	2.68	67
1706	1.75	3.49	5.32	10.45	900.1	0.59	27.2	203	2.58	66
1633	4.05	3.43	5.1	12.62	875.4	0.71	25.3	53	3.49	69
0894	2.44	3.48	5.07	10.96	907.1	0.67	26.4	137	2.67	63
1703	2.52	3.22	5.27	10.97	854.5	0.63	21.5	37	0	0
1458	3.66	3.36	4.92	12	947.3	0.72	25.2	65	2.43	65
1619	1.61	3.35	5.21	10.05	914.8	0.62	26.1	82	2.89	70
1640	1.03	3.1	5.41	9.43	899.6	0.59	22.4	230	2.59	68
1634	3.53	3.34	5.11	11.96	885.1	0.69	27.7	270	2.61	66
1017	2.48	3.04	5.17	10.65	926.1	0.68	24.2	310	2.28	65
1659	3.79	2.75	4.82	11.48	935.1	0.72	24.5	126	0	0
1656	1.57	2.82	5.25	9.64	910.5	0.66	27.7	70	0	0
1714	2.06	3.05	5.29	10.35	870.2	0.63	22.8	25	0	0
1709	1.19	3.1	5.44	9.61	858.2	0.57	25.6	21	0	0
14022	2.93	3.42	4.94	11.25	978.1	0.7	28.8	94	3.28	69

图1 数据样本

然后，结合牛只的其它信息，如出生日期、胎次、泌乳天数等，对数据进行综合分析。

```
def result_jcjg_jisuan():
    connection = pymysql.connect(host = local,port = 8888,user = usr,password = pwd,db = mydb,charset = 'utf8')
    result_jcjg=pandas.read_sql('select CATTLE_NUMBER,BULL_ID,SAMPLING_DATE,STANDARD_EAR_NUMBER,FETAL_TIMES,DATE
    result_jcjg.columns=list({"牛场编号","牛号","采样日期","标准耳号","胎次","产转日期","产转间隔","泌乳天数",
    result_history=pandas.read_sql('select  CATTLE_NUMBER,BULL_ID,SAMPLING_DATE,STANDARD_EAR_NUMBER,FETAL_TIMES,
    result_history.columns=list({"牛场编号","牛号","采样日期","标准耳号","胎次","产转日期","产转间隔","泌乳天数",
    connection.close()

    result_jcjg['隐重比'] = result_jcjg.apply(lambda x:x["重脂率"]/x["重白率"],axis=1)#计算脂重比
```

Python代码1

最后，将分析结果直接生成word版分析报告，为奶牛场提供奶牛养殖方面参考。利用Python直接生成word文档推荐使用mailmerge包，可实现文档内简单表格、多表头表格的生成，表格内插入多条数据，插入matplotlib画图，以及文档名称插入变量和生成时间等信息。

```
template= "D://母版.docx"#路径
document=MailMerge(template)
document.merge(采样日期=str(result_jcjg["采样日期"][1])[0:10],牧场名称=result_j
biao1_word=json.loads(biao1.to_json(orient="records",force_ascii=False))
document.merge_rows("牛号",biao26_word)
document.write("牛场编号"+result_jcjg["牛场编号"][1]+"_分析报告_时间"+str(ti

if __name__ == "__main__":
    result_jcjg_jisuan()
```

Python代码2

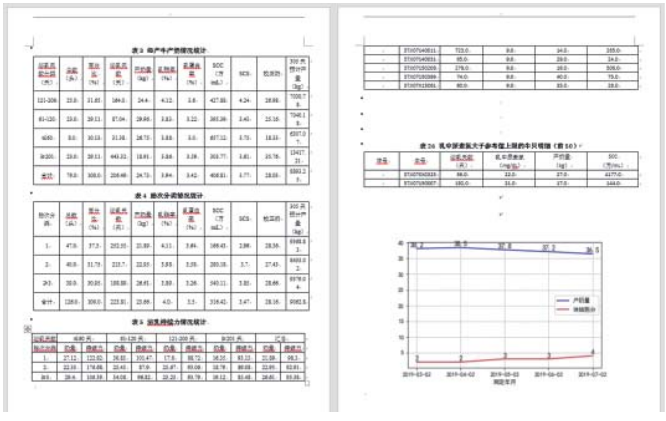


图2 Word报告

2 奶牛选种选配

奶牛场的生产水平不仅与养殖技术、养殖环境等硬件环境相关，更重要的是奶牛本身的产奶性能，比如在相同条件下，产奶性能高的牛能产更多的牛奶，进而给奶牛场带来更大

的经济效益。所以，要提高整个奶牛场的效益，选种选配计划至关重要，我们要做的就是用数据做一份最佳的选种选配方案，避免人为主观影响。

首先，做选种选配计划需要的数据有奶牛3-6代的系谱和奶牛育种值数据。

然后，奶牛场内参与选种选配计划的牛只数量有限（1000头左右），数据量不算大。

```
df2 <- data.frame(ind=inds,father=ff,mother=mm)
df2=df2[complete.cases(df2),]

xipu_he2 <- rbind(xipu,df2)
xipu_check2 <- prePed(xipu_he2)#检验系谱
inbreding2 <- pedInbreeding(xipu_check2)#计算近交系数
colnames(inbreding2) <- c("ind","inbreeding")#修改列名
result2 <- merge(inbreding2,df2,by="ind",all.y = T)
result2 <- result2[,2:4]#近交系数最终结果
inbreding_cpi_value <- merge(result2,cpi,by.x="father",by.y = "BULL_ID",all.x = T)
inbreding_cpi_value2 <- merge(inbreding_cpi_value,cpi,by.x="mother",by.y = "BULL_ID",all.x = T)

inbreding_cpi_value2["CPI_progeny"]=(inbreding_cpi_value2["CPI.x"]+inbreding_cpi_value2["CPI.y"])/2
inbreding_cpi_value3 <- inbreding_cpi_value2[,c("mother","father","inbreeding","CPI_progeny")]
inbreding_cpi_value3 <- inbreding_cpi_value3[order(-inbreding_cpi_value3["CPI_progeny"],inbreding_cp
```

R语言代码

遍历计算每头种公牛和每头待配母牛的近交系数和后代育种值，计算结果结果按优先顺序排序，生成一个Excel报告给奶牛场提供选种选配参考。此项目利用R语言完成，optiSel包可实现种公牛和待配母牛的近交系数的计算。

案例二：企业舆情监督

公安部门对敏感企业进行舆情监督，使用Python爬虫定期获取企业舆情数据，进而对数据进行情感分类，得到好评率、媒体指数等企业打分指标。下面以企业评论数据的情感分类为例，介绍实现过程。

1　加载情感词典

从数据库读取企业爬虫数据，同时读取已有的情感词典（包括正向情感词和负向情感词）、否定词词典和程度副词词典作为全局变量。

```python
def read():
    # 读取停用词文件
    fr = codecs.open('stopwords.txt', 'r', encoding='utf-8')
    for word in fr:
        stopwords.add(word.strip())
    fr.close()

    # 读取情感字典文件
    sen_file = open('BosonNLP_sentiment_score.txt', 'r+', encoding='utf-8')
    # 获取字典文件内容
    sen_list = sen_file.readlines()
    # 读取字典文件每一行内容，将其转换为字典对象，key为情感词，value为对应的分值
    for s in sen_list:
        # 每一行内容根据空格分割，索引0是情感词，索引01是情感分值
        if len(s.split(' '))==2:
            sen_dict[s.split(' ')[0]] = s.split(' ')[1]
    sen_file.close()

    not_word_list=[]
    #读取否定词文件
    not_word_file = open('notDic.txt', 'r+', encoding='utf-8')
    # 由于否定词只有词，没有分值，使用list即可
    for i in not_word_file.readlines():
        not_word_list.append(i.split('\n')[0])
    not_word_file.close()
```

Python代码3

2 文本分词

将文本利用结巴分词进行处理并去除停用词，得到文本的单词列表。

```python
def seg_word(sentence):
    #使用jieba对文档分词
    seg_list = jieba.cut(sentence)
    seg_result = []
    for w in seg_list:
        seg_result.append(w)
    # 去除停用词

    return list(filter(lambda x: x not in stopwords, seg_result))
```

Python代码4

3 生成文本词典

将单词列表中的词归类，生成此文本的情感词词典、否定词词典和程度副词词典，其key值为此单词在文本中的位置索引，value值为词的得分。

```
def classify_words(word_list,not_word_list):

    #词语分类,找出情感词、否定词、程度副词
    # 分类结果，词语的index作为key,词语的分值作为value，否定词分值设为-1
    sen_word = dict()
    not_word = dict()
    degree_word = dict()
    num=0

    # 分类
    for index in range(0,len(word_list)):
        if word_list[index] in sen_dict.keys() and word_list[index] not in not_word_list and word_list[index] not in degree_dic.
keys():
            # 找出分词结果中在情感字典中的词
            sen_word[index] = sen_dict[word_list[index]]
            num=num+1
        elif word_list[index] in not_word_list and word_list[index] not in degree_dic.keys():
            # 分词结果中在否定词列表中的词
            not_word[index] = -1
        elif word_list[index] in degree_dic.keys():
            # 分词结果中在程度副词中的词
            degree_word[index] = degree_dic[word_list[index]]
    # 将分类结果返回
    return sen_word, not_word, degree_word,num
```

Python代码5

4 计算情感极性得分

两个情感词之间的否定词和程度副词与其中后一个情感词构成一个情感词组，所有情感词组的得分之和即为文本的情感极性得分。

公式如下：

$$\text{score} = \sum_{i} (-1)^{a_i} \times b_i \times c_i$$

其中a_i为第i个情感词组中的否定词词数，b_i为此词组中所有程度副词的权值之积，c_i为情感副词的得分。

```
while(i<len(seg_result)):
    if i in sen_word.keys():
        #score1=W*float(sen_word[i])   #这个也有影响！！！
        #score+=score1
        score+=W*float(sen_word[i])
        W=1
        sentiment_index += 1
        if sentiment_index < len(sentiment_index_list) - 1:
            # 判断当前的情感词与下一个情感词之间是否有程度副词或否定词
            for j in range(sentiment_index_list[sentiment_index], sentiment_index_list
                # 更新权重，如果有否定词，取反
                if j in not_word.keys():
                    W *= -1
                elif j in degree_word.keys():
                    # 更新权重，如果有程度副词，分值乘以程度副词的程度分值
                    W *= float(degree_word[j])
```

Python代码6

5 得分情况展示

得分	文本
12.31096	原先是 XX4X，屏碎了，换了 XX7。XX7 缺点：无单独 TF 卡，4G 上网信号不好，电池续航不够，优点：相比 4X 内存大，速度快
12.08848	觉得 RR 现在是中国之光，国之骄傲的点个赞！RR 的这个网站方便我准点抢手机，比手机端快一点哈哈
11.92762	今天来 RR 商城处理了三部手机，感觉不错。支持 RR 商城继续做以旧换新业务，还会继续支持这项公益事业的。
11.78427	就是喜欢 RR 的产品！嘿嘿嘿嘿！
11.61348	很靠谱的产品！支持 RR！希望 RR 能一直不骄不躁地走下去！
0.023507	虚假宣传，偷工减料，中央空调主机不放设备空间，放在阳台占使用面积；
0	这
-0.05553	PP 手机电一点都不经用　今天早上刚充满的电　什么没有用　电都会少太烦
-0.06496	刚刚更新完不仅主题没了，一按返回功能也被取消现在用下面的隐藏键十分不方便！强烈要求 RR 恢复此功能。。也不说清楚感觉被骗更新差评。失望
-0.0657	MM 手机太烂不要买，买了就上当了，OO 不一样

表1 打分结果

6 确定分类范围

利用文本得分划分文本情感正负向性或者中立性。

（1）观察得分，发现0分并不是合理的正负向分界线，于是将分类问题抽象成最优化问题，即寻找最优的中立分数的上下限，使所得的分类与已知分类相比正确率最高。而得到这个范围之后，即可应用到其他文本的分类标准。

（2）可行域根据样本分数确定，如根据排序后分数合理百分比的中间段数据的极差确定，此处下界可行域为(-2,4)，上界可行域为(-1,6)。目标函数为分类正确率。如果新上下界的正确率高于旧上下界，则更新上下界。

（3）得到中立上界为3.7分，中立下界为-1分，正确率为86.24%。

7 不足之处

（1）分数绝对值低的部分样本判断准确性不算高，分类不够明确。

（2）分析文本情感色彩时没有考虑语境和句间关系，容易误判如讽刺性文本等特殊文本。

（3）对情感词典有一定依赖性，后期可根据不同平台上的样本利用机器学习的朴素贝叶斯算法对情感词典进行优化，使之更适用于特定平台上的文本情感判断。

联系方式：山东智谷数据分析师事务所
联系人:苏航
联系电话:0538-8932988
办公地址:山东省泰安高新区南天门大街1110号

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：王丹 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