

分析篇

第 1 章

数据分析概述

数据分析作为信息技术的重要组成部分,已经成为许多领域不可或缺的关键工具。它不仅能够帮助我们更好地理解复杂的现象,还能够挖掘出隐藏在数据背后的有价值信息,为决策提供强有力的支持。本章首先介绍数据分析的概念及类型;其次介绍数据分析的思维框架、流程框架和常见的误区;然后介绍数据分析的方法论以及常用的数据分析方法;最后介绍数据分析在现实领域的应用。

本章学习目标

- (1) 掌握数据分析的概念和类型;
- (2) 理解数据分析的思维框架;
- (3) 掌握数据分析的流程框架;
- (4) 掌握数据分析的方法论;
- (5) 掌握常用的数据分析方法;
- (6) 了解数据分析在现实领域的应用情况。

1.1 数据分析的概念及类型

1.1.1 数据分析的概念

数据分析可以追溯到古代文明时期,人们通过观察天气、温度和季节的变化来决定何时种植庄稼,这可以看作最早的一种数据分析形式。18 世纪正态分布曲线的提出为数据分析奠定了基础,20 世纪计算机技术的出现和不断进步使数据分析得以快速发展,而大数据时代机器学习(machine learning)、深度学习(deep learning)等人工智能技术的兴起则进一步推动了数据分析的发展和应用。

数据分析可以视为数学与计算机科学相结合的产物,指的是采用适当的统计分析方法和工具对收集来的大量数据进行系统化处理、理解和解释,旨在从原始数据中发现有价值的信息、洞察潜在的规律、揭示隐藏的趋势与关联,并通过可视化和其他形式的报告将这些发现以易于理解的方式呈现出来,以支持决策的制定。简言之,数据分析就是为了提取有用信息和形成结论而对数据进行详细研究与概括总结,它的目的在于最大化地开发数据的功能,充分发挥数据的作用。

从数据分析的概念中不难看出,数据分析涉及三个核心问题,分别为数据、分析过程和分析结果。其中,数据是数据分析的基础,分析过程直接关系到是否能够得到预期的、正确的分析结论,而分析结果则是定位问题和解决问题的依据。

1.1.2 数据分析的类型

数据分析可以按不同的维度进行划分。

1. 按分析目的划分

按分析目的的不同,数据分析可分为描述性分析(descriptive analytics)、探索性分析(exploratory analytics)、预测性分析(predictive analytics)和规范性分析(prescriptive analytics)。

(1) 描述性分析: 数据分析中最为基础和常见的分析类型。其主要是对已经发生的事件进行描述和总结,目的是通过一系列统计量来描述数据的基本特征和规律,回答“已经发生了什么?”。

(2) 探索性分析: 通过作图、制表、方程拟合、计算特征量等手段,在尽量少的先验假设下探索数据的结构和规律,其目的是发现数据中的因果关系,以及检测可能存在的异常值或错误,回答“为什么发生?”。

(3) 预测性分析: 通过建模、统计分析、机器学习等手段,基于历史数据和其他相关信息,对未来趋势或结果进行预测,其目的是通过挖掘数据中的模式和关联,发现数据中的潜在价值,为决策提供支持,回答“将发生什么?”。

(4) 规范性分析: 在描述性分析和预测性分析的基础上,通过优化算法和决策模型,提出具体的行动建议或方案,其目的在于通过深入研究和评估现有的情况或问题,为决策者提供最优的行动指南,回答“如何优化即将要发生的?”。

2. 按分析方法划分

按分析方法的不同,数据分析可分为定性分析(qualitative analytics)、定量分析(quantitative analytics)和混合方法分析(mixed analytics)。

(1) 定性分析: 对数据的性质、特征、发展变化规律进行描述和解释,主要依赖于非数值化的信息,如文档资料、访谈记录、观察笔记等,它可以提供对数据深入而丰富的描述,揭示出隐藏在表面现象背后的复杂性和多样性。

(2) 定量分析: 通过统计学方法和数学模型对数据进行分析与处理,主要依赖于数值化的数据,如经济指标、金融数据、生物医学数据、商业数据等,它可以揭示数据的数量特征、数量关系和数量变化。

(3) 混合方法分析: 将定量分析和定性分析相结合,综合运用多种方法和技术进行数据分析。

3. 按分析数据的类型划分

按分析数据的类型不同,数据分析可分为结构化数据分析(structured data analytics)、非结构化数据分析(unstructured data analytics)和半结构化数据分析(semi-structured data analytics)。

(1) 结构化数据分析：该类分析针对的数据类型为结构化数据库中的数据，如关系型数据库中的表格数据。

(2) 非结构化数据分析：该类分析针对的数据类型为数据结构不规则、不完整，没有预定义的数据模型，如文本、图片、音视频、XML(可扩展标记语言)文件等。

(3) 半结构化数据分析：该类分析针对的数据类型为包含在两个或多个数据库(这些数据库含有不同模式的相似数据)中的数据，这些数据具有一定的结构，但结构的变化很大，如 OEM(object exchange model)就是一种典型的半结构化数据模型。

总的来说，数据分析的分类取决于分析的目的、方法、数据类型等多个方面，不同的分类方式是为了更好地理解和应用数据分析技术，以满足不同领域和场景的需求。

1.2 数据分析框架

1.2.1 思维框架

数据分析的思维框架是指导数据分析过程的系统化思考方式，能够帮助数据分析者在面对复杂数据时保持清晰、有条理的思维逻辑。数据分析思维框架强调以明确的目标为导向，通过一系列有序的步骤和方法，从数据的收集、整理、探索到分析和解释，最终实现对数据的深入理解和有价值的信息提取，具体如图 1-1 所示。

1. 为什么要分析

这个部分主要是明确分析的目的，对于确保后续分析过程聚焦在关键问题上至关重要。例如，从业务需求的角度可以是为了解决某个具体问题、优化某个流程或是实现特定的业务增长；从用户或客户需求角度可以是为了更好地了解用户或客户的行为、偏好和需求；从市场和竞争环境角度可以是为了分析市场趋势或竞争对手的情况、了解行业内的变化和潜在机会。

2. 分析什么

这个部分主要是明确分析的对象，确保分析工作与组织的整体战略和目标保持一致，为后续的分析工作提供明确的方向和重点。例如，解决业务问题需要分析与业务直接相关的关键指标数据；了解用户行为需要分析用户在产品或服务中的点击、浏览、搜索、购买等行为数据；了解竞争对手需要分析竞争对手的产品、市场策略和业绩表现等数据。

3. 怎么分析

这个部分涵盖的内容较为宽泛，也是思维框架中的核心环节。例如，当数据集较大时，如何选择合适的抽样方法，确保样本的代表性和可靠性；如何根据数据分析的目标，选择合

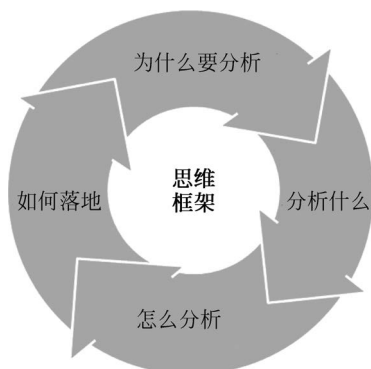


图 1-1 数据分析的思维框架图

适的分析模型和方法,如描述性统计、预测建模、机器学习等;考虑使用哪些技术或工具来辅助分析,如 Python 或 R 语言、可视化工具等;如何设计恰当的统计测试或实验来验证前期提出的相关假设。

扩展阅读 1-1 构建数据安全保护和政策沟通协调机制



4. 如何落地

这个部分需要综合考虑多个方面,包括业务目标、业务流程、策略制定和执行力等,是确保分析结果转化为实际业务价值的关键步骤。例如,与业务部门共同制订实施计划,将分析结果转化为具体的业务决策或行动建议;建立跨部门的数据分析团队,加强团队间的沟通与协作,确保分析工作的顺利推进;通过制度和流程建设,确保数据分析工作的规范化和可持续性;跟踪和评估实施效果,及时调整分析策略,确保分析结果的持续价值。

1.2.2 流程框架

数据分析是一个动态且复杂的过程,涉及多个环节和技术,为了保证数据分析工作的规范性和有效性,构建一个科学、系统的数据分析流程框架显得尤为重要,不仅能够提高分析效率,还能够确保分析结果的准确性和可靠性。一般来说,数据分析的流程框架可以划分为六个阶段,分别为目标解析、数据采集、数据处理、数据分析、结果呈现和报告撰写。

1. 目标解析

目标解析是数据分析的起点。每次数据分析任务开始之前,首先需要弄清楚这次分析是为了解决什么问题,在确定总体分析目标之后,再通过拆解的方式将总目标细化分解为多个子目标,明确先分析什么、后分析什么,理顺各个子目标之间的逻辑关联性,搭建数据分析的逻辑框架。只有带着清晰的目标进行数据分析,才能为后续的各个步骤提供正确的指引方向,确保数据分析过程的有效推进。

2. 数据采集

数据是分析开展的基础,为数据分析提供直接的素材和依据。但是数据并不会凭空出现,需要根据分析目标,选择合适的数据源采集分析所需的数据。数据获取的方式主要有两种:直接获取和间接获取,前者获取的为一手数据,后者获取的为二手数据。直接获取主要是通过统计调查或者科学实验的方法得到第一手或直接的统计数据,间接获取则是通过查阅资料、数据统计工具等获取数据,这类数据往往来源于他人的调查或实验,是经过加工整理后的数据。常见的数据获取渠道有数据库、互联网、公开出版物、数据统计工具、市场调查等。

3. 数据处理

数据处理是数据分析前至关重要的一步。一方面,数据采集所获取的原始数据可能会包含大量的噪声、错误和冗余信息,需要通过数据处理,消除不利因素,提高数据质量,保证后续分析能够在准确和有效的数据基础上进行;另一方面,不同的数据分析方法和工具对数据的格式与类型可能有不同的要求,需要通过数据清洗(data cleaning)、转换和整合等处

理步骤将数据转化成符合分析需求的格式,保证分析过程的顺畅和高效。对数据的处理方式可以是将相似的数据进行归类整理,或是将有关系的数据组织起来,或是对整理好的数据进行算术和逻辑运算,或是对数据进行排序等。

4. 数据分析

数据分析是整个流程中的重中之重,涉及统计学、机器学习、数据挖掘等技术与方法的综合运用。进行数据分析时,需要根据分析目的选择合适的分析工具和方法,发现数据中的模式、趋势和关联,提取有价值的信息。数据分析遵循的思维方式通常有分类思维、回归思维、假设思维、对比思维、相似匹配思维等。常用的数据分析方法有对比分析法、分组分析法、结构分析法、交叉分析法、漏斗分析法、矩阵分析法、综合评价分析法等。在进行数据分析的过程中,需要评估分析数据是否完整、真实且有效,是否对数据进行了深入的剖析,是否能够获得准确的数据分析结论。

5. 结果呈现

结果呈现阶段的主要任务是以清晰、直观的方式展示分析结果,确保相关人员充分理解和应用结果。在进行结果呈现时,一方面要根据分析目的和受众特点,选择恰当的图表类型来展示数据,在保证直观的同时突出关键结论,方便受众更好地把握数据特征和规律;另一方面也需要充分考虑受众的需求和背景知识,避免使用过于专业或复杂的术语,保证受众能够轻松理解分析结果,此外还要注意避免信息过载和复杂设计,保持呈现内容的简洁、明了。

6. 报告撰写

报告撰写是数据分析的收尾环节,不仅仅是对所有分析工作的系统总结和精准展现,更是实现数据价值转化的关键桥梁。通过撰写报告,能够将复杂的数据分析结果以清晰、有条理的方式呈现出来,为决策者提供清晰、有用的信息和建议,帮助他们作出科学、正确、严谨的决策。在内容方面,报告中应包含数据分析的起因、过程、结论和建议;在结构方面,报告应做到结构清晰、主次分析,具有一定的逻辑性,对每一个问题都给出明确的结论,一个分析对应一个结论;在可读性方面,报告中不要出现过多的专业术语,尽量按照受众的需求通过图表化的方式呈现出来。

1.3 数据分析常见误区

1.3.1 盲目采集数据

数据分析人员要避免盲目地追求数据的全面性。诚然,增加数据采集量在一定程度上对分析工作具有积极的影响,但是这种影响是存在局限性的。盲目采集数据可能会带来如下问题。

(1) 资源浪费:大量采集无关数据不仅会消耗存储空间,增加数据传输和处理的成本,降低系统性能,还可能导致关键数据被无关数据淹没。

扩展阅读 1-2 提
升数据质量,完善
数字政府体系框架



(2) 数据质量低:盲目采集数据可能会引入一些无关或误导性的信息,如错误的数据、重复的数据、不完整的数据或格式不一致的数据等,进而影响分析结果的准确性和有效性。

(3) 分析效率低:数据分析人员需要花费更多的时间和精力来清洗与处理海量无关的数据,将会大大降低分析的效率。

(4) 结果不准确:如果分析所基于的数据是无关或错误的,分析结果可能是不准确的,甚至可能是错误的,进而导致错误的决策和行动。

1.3.2 脱离实际业务

数据分析人员虽然具备强大的数据处理和分析能力,但由于缺乏实际业务经验,他们所做的分析可能与实际业务逻辑脱节,造成分析结论的不全面和不准确。脱离业务实际可能会带来如下问题。

(1) 分析目标不明确:脱离业务实际进行数据分析,可能导致分析目标模糊,无法为业务决策提供有针对性、有价值的洞见,削弱数据分析的实际意义和应用价值。

(2) 数据解读错误:分析人员如果缺乏对业务背景的深入了解,可能难以精准解读数据的内在含义,甚至陷入误解,导致得出错误的结论或提出不切实际的建议。

(3) 难以落地实施:即使数据分析的结果是准确的,但由于一开始与实际业务场景存在较大的偏差或不符,那么分析结果也将难以在实际业务操作中落地实施,不能转化为切实的业务成果,数据的潜在价值也无法充分发挥。

1.3.3 分析不够深入

数据在收集和整理之后需要经过深入的分析,才可能获得有价值的信息,如果只关注数据的表面特征,将很难为后续工作提供实质性的帮助和推动。缺乏深入的分析可能会带来如下问题。

(1) 难以发现问题:有时表面数据可能掩盖了实际情况中的复杂性和不确定性,不经过深入分析将很难发现那些隐藏在数据中的微妙变化和潜在危机,导致不能及时发现潜在问题或风险。

(2) 错误决策:数据分析不深入容易造成对数据的片面理解,而忽视数据背后的深层次信息和潜在趋势,从而导致错误的决策,给组织带来不必要的风险和偏差。

(3) 丧失竞争优势:深入剖析数据可以帮助企业洞察市场趋势、理解消费者需求、优化产品策略,以及精准制订营销方案等,助力企业在激烈的市场竞争中脱颖而出;否则,可能导致企业错失很多宝贵的信息和机会,逐渐失去竞争优势,甚至面临被市场淘汰的风险。

1.4 数据分析的方法论

做好数据分析工作,不仅仅是要掌握数据分析的各种工具和方法,更需要懂得数据分

析原理。缺乏理论的指导,就如同盲人摸象,既不知道应从哪些关键点切入分析,也难以确定分析的重点。数据分析的方法论主要是从宏观角度为数据分析工作提供系统性的指导,更像是数据分析前的一个总体规划,为后期具体数据分析工作指明方向,确保数据分析工作有条不紊地进行。

1.4.1 5W2H

5W2H方法又称为七问法,这种方法简单方便,容易理解和使用,任何事情都可以通过这七个方面去思考,数据分析同样也不例外,它可以帮助我们全面、系统地思考数据分析的各个方面。5W2H是5个以W和2个以H开头的英文单词缩写,具体含义如下。

(1) why(何故?)——为什么要做数据分析?是因为遇到了问题,还是为了优化某个流程?回答该问题是为了更好地确定分析的重点和方向。

(2) what(何事?)——要分析的是什么?想得到什么结果或结论?回答该问题是为了更好地确定分析对象,定位分析目标。

(3) who(何人?)——谁会参与此次分析?是分析人员、业务人员还是其他相关人员?回答该问题有助于合理分配任务和资源,确保分析的顺利进行。

(4) when(何时?)——分析的时间范围是什么?是过去的还是实时的?回答该问题有助于明确分析的时间节点,确保分析的及时性和准确性。

(5) where(何地?)——数据从哪里来?内部数据库、第三方平台或其他渠道?回答该问题有助于评估数据的可靠性和有效性。

(6) how(何法?)——如何实施分析?使用什么工具和方法?如何清洗、处理和可视化数据?回答该问题能保证分析按照既定流程进行,提高分析的效率。

(7) how much(多少?)——分析到什么程度?分析需要多少人力、时间、成本等资源?回答该问题是为了确定问题要解决到什么程度,以及评估分析方案的投入产出比,确保在预算和既定时间内完成分析任务。

1.4.2 逻辑树

逻辑树是麦肯锡公司提出的一种重要的分析问题、解决问题的方法,又称问题树、演绎树或分解树。该方法将已知问题视为“树干”,将与问题相关的子问题视为“树枝”,大“树枝”上还可以继续延伸出更小的“树枝”,以此类推,找出所有和已知问题相关联的问题,将一个复杂的问题分解成一系列便于操作的子问题。这种方法既有助于厘清思路,避免重复和无关的思考,保证问题解决过程的完整性,也便于确定解决各子问题的先后顺序,将责任落实到个人。

逻辑树需要满足如下三个要素。

(1) 要素化:将相同问题总结归纳成要素,并从中识别出关键要素。

(2) 框架化:遵循不重复、不遗漏的原则,将各个要素组织成框架,即各要素虽然在逻辑上独立,但组合起来后应该能够全面覆盖问题。

(3) 关联化:框架内各要素不能孤立存在,要保持必要的相关性,共同构成一个完整的

逻辑体系。

图 1-2 为逻辑树在“某品牌手机线上销售额下降”分析中的应用示意图。



图 1-2 逻辑树在“某品牌手机线上销售额下降”分析中的应用示意图

1.4.3 多维度拆解

多维度拆解方法侧重于从多个角度或维度来考察问题,其关键在于识别分析问题所需的时间、空间、社会经济因素、技术因素等,一旦确定分析维度,数据分析人员就可以从这些维度拆解问题,分析每一个维度对问题的影响。以“提升某品牌手机的市场份额”为例,可以从产品特性、竞争对手、销售渠道等维度展开分析。

- (1) 产品特性: 自身品牌手机的功能、质量、价格等因素是否符合市场需求。
- (2) 竞争对手: 竞争对手品牌手机的特性、销售策略、市场份额等。
- (3) 销售渠道: 不同销售渠道的销售效果的对比,是否存在优化和拓展空间。
-

使用多维度拆解方法的一个重要环节就是明确从哪些维度去进行拆解,一般可以从指标构成或者是业务流程的角度进行拆解。如销售额下降问题,从指标构成角度进行拆解可以进一步明晰与销售额下降直接相关的指标,从销售流程角度进行拆解可以找到导致销售额下降的关键业务节点。多维度拆解的思维方式并非一成不变,同一个问题的拆解方式也

有很多种,但万变不离其宗,无论如何拆解,都是为了更好地完成数据分析工作,实现问题的解决或优化。

1.5 常用的数据分析方法

相较于数据分析的方法论而言,数据分析方法就是从微观的角度去指导数据分析工作的开展。常用的数据分析方法多种多样,本节主要介绍如下七种:描述性分析法、对比分析法、相关性分析法、群组分析法、假设检验分析法、归因分析法和预测分析法。

1.5.1 描述性分析法

描述性分析法一般用于回答“是什么”及“怎么样”的问题,它既是最直观的数据分析手段,也是数据分析最基础的工作。但该方法侧重于呈现现状,并不会解释“为什么”,因此一般用在需要对数据进行初步了解、特征描述、异常检测和可视化展示时。描述性分析法通常从以下指标维度展开。

1. 集中趋势指标

- (1) 均值:数据的平均数,反映的是数据整体的平均水平。
- (2) 众数:数据集中出现次数最多的数值,反映的是数据的集中点。
- (3) 中位数:将数据按照从小到大的顺序排列后,位于中间位置的数值。在对称分布的数据集中,均值和中位数较为接近;在偏态分布的数据集中,均值和中位数的偏差较大。中位数最显著的特点在于它不易受极端值的影响,更为稳健。

2. 分散程度指标

- (1) 方差和标准差:方差是数据集中各数据与均值的差的平方和的平均数,标准差是方差的算数平方根,它们反映的是数据的离散程度,值越大,说明数据集中数值的波动越大。
- (2) 四分位数:将数据按照从小到大的顺序排列后分成四个等份,每份包含数量相同的数据点,处在三个分割点位置的数值分别为下四分位数(Q1,25%位置的数值)、中位数(Q2,50%位置的数值)和上四分位数(Q3,75%位置的数值)。该类指标可以反映数据的整体分布情况,也可以识别出可能的异常值。

3. 偏态与峰态指标

- (1) 偏态系数:用于描述数据分布对称性的指标。偏态系数为0时,表示分布是对称的;偏态系数大于0时,表示分布是右偏的;偏态系数小于0时,表示分布是左偏的。
- (2) 峰度系数:用于描述数据分布峰顶陡峭或扁平程度的指标。峰度系数为0时,表示分布与正态分布基本一致;峰度系数大于0时,表示分布呈尖峰状;峰度系数小于0时,表示分布呈扁平状。

4. 其他指标

- (1) 极差:数据集中最大值与最小值之差,反映的是数据波动的范围。该指标对异常

值较为敏感。

(2) 标准分(z 分数): 原始数据与均值之差除以标准差的商,反映的是一个数据点相对于均值的位置。该指标不会受到原始测量单位的影响。

1.5.2 对比分析法

对比分析法是日常生活中最常见也最常用的一种方法。这种方法就是将两个或两个以上的数据放在一起进行比较,分析它们之间的异同,从中发现数据蕴含的有价值的信息,进而作出正确的判断和评价。对比分析法可以分为横向对比和纵向对比。

1. 横向对比

横向对比是指将两个或多个同类事物放在同一层面上进行比较,具体可以有以下一些比较维度。

(1) 与既定目标比: 将实际完成情况与预设目标进行比较。如电商在“双十一”之前都会有自己预定的销售业绩,可以将实际销售业绩和预定销售业绩进行对比,看看是否完成了既定的目标,对未完成的原因进行分析。

(2) 与行业内比: 与行业内的标杆、竞争对手或平均水平进行比较。如某品牌手机生产商将自身的产品参数与行业内“领头羊”的产品参数进行对比,寻找改进的空间。

(3) 与同级部门、单位、地区比: 这种比较可以了解自身某方面或各方面的发展处于什么样的水平。如某高校将自身的师资力量、办学水平等与同级别的其他高校进行对比,明晰自身的优劣势,确定今后的发展方向。

2. 纵向对比

纵向对比是指对同一事物在不同时间或不同条件下的状态进行比较,具体可以有以下比较维度。

(1) 与不同时期比: 对同一指标在不同时间窗口下的数值进行比较。如电商将本年“双十一”的销售业绩与前两年同期的进行对比,对增长或下降的原因进行分析,为来年的销售做准备。

(2) “事发”前后对比: 对某件“事情”发生前后的数据进行比较。如某手机销售商为扩大手机的销售量制订了一套促销方案,对促销方案实施前后的销售量进行对比,挖掘对销售有促进作用的关键要素。

使用对比分析法需要注意的是,要保证对比对象之间是存在可比性的,同时对比的指标类型必须是一致的,而且数据应该具有相同的计量单位。

1.5.3 相关性分析法

相关性分析法也是日常生活中较为多见的一种数据分析方法,这种方法主要用于揭示两种或两种以上的数据之间是否存在某种依存关系,如人的身高和体重之间、空气的相对湿度和降雨量之间、工作压力和死亡率之间等。两种数据之间如果存在依存关系,则视为

有相关性；否则没有。这种相关性一般使用“相关系数”来衡量，相关系数的正负可以反映两者之间的相关性是同向的还是反向的，绝对值的大小则可以反映相关程度的高低。虽然对相关程度的判断并没有统一标准，但通常按相关系数绝对值大小分成三个区间：0~0.3 被视为低度相关，0.3~0.6 被视为中度相关，0.6~1.0 被视为高度相关。由此可见，相关系数的绝对值越大，表明两种数据之间的相关程度越高。

相关性分析法除了能够对既定数据之间的相关性进行分析，还能够帮助分析人员拓展思路，如在对“销售额下降”问题进行分析时，可以先通过相关性分析法找到对销售额具有影响的因素，再根据相关系数的大小快速锁定主要因素，进而进行有针对性的分析。在众多数据分析方法中，相关性分析法属于比较通俗易懂的，即使没有关于分析方法背景知识的受众，也比较容易理解和接受相关性分析的结果，这也是相关性分析法应用多且广的重要原因之一。

虽然相关性分析法有很多优势，使用也较为方便，但是如果使用不当，也可能会导致错误的分析结果。特别需要注意的是：相关关系并不是因果关系。例如：有调查数据显示冰激凌的销量和溺水事故发生的次数存在相关性，但是这并不意味着冰激凌销量的增加是导致溺水事故增多的原因，两者不具备因果关系，两者之间的这种相关性主要是因为夏天温度升高，人们更倾向于吃冰激凌和更多地参与水上运动，而后者才是导致溺水事故增加的原因。由此可见，在进行相关性分析时，对待相关关系和因果关系的分析要更为谨慎，避免得出错误的因果结论。

1.5.4 群组分析法

群组分析法常用于研究不同群体的行为模式和动态变化，其核心是根据分析目的，按照某个特定的特征，对数据中具有不同特性的对象进行区分，把有相同特性的数据合并在一起，通过各组数据的比较，发现其中的联系和规律，所以该方法通常需要配合对比分析法一起使用。

分组数量是使用群组分析法时需要特别关注的问题，分组数量的多少直接关系到分析的细致程度和结果的准确性，分组过少会造成数据分布集中，无法充分揭示数据的特征和差异，分组过多会导致数据过于分散，结果过于复杂和难以理解，两种情况都不利于观察数据的分布特征和规律，有碍数据分析工作的开展，因此在分析之前需要确定好合适的分组数量。分组通常可以参照如下标准。

(1) 人口统计学特征：如性别、年龄、职业等，这种分组标准通常可以用于市场研究或科学研究中，以了解不同人群的行为或偏好。如某城市想要分析不同年龄段人口的就业情况，以便制定更有效的就业政策，可以按照合适的年龄段区间对数据进行划分。

(2) 时间序列：将数据按照时间段进行分组，如按照月份、季度或年份进行划分，这种分组标准通常用于趋势和周期性模式的分析。如某电商平台想了解过去一年内销售额的变化情况，以制定更精准的营销策略，平台收集了每日的销售额数据，并按月份对数据进行分组。

(3) 业务场景：根据企业或行业的不同业务场景对数据进行分组，这种分组标准有助于更深入理解各业务场景的特点，为制定更有效的决策提供支持。如某视频平台想要考察

平台的用户留存率,可以根据用户的注册月份对用户进行分组,以观察用户留存率与月份之间是否存在关联性。

1.5.5 假设检验分析法

假设检验分析法也是数据分析方法中非常重要的一种,主要用于根据样本数据对总体参数或总体分布形式作出推断,其核心思想是先提出关于总体参数的假设,再通过样本数据来检验这个假设是否成立,从而为决策者作出科学、客观的决策提供支持。该方法主要包括四个步骤。

1. 提出假设

根据要解决的问题提出研究假设。例如,某糖果生产商想了解新的糖果配方是否比传统配方更受小朋友的欢迎,可以提出如下假设。

H_0 (原假设): 新配方和传统配方在小朋友中的受欢迎程度没有差别。

H_1 (备择假设): 新配方比传统配方更受小朋友欢迎。

2. 收集数据

对于之前提出的假设,不能靠主观猜想或臆断来证明假设的成立与否,而是需要找到证据证实假设成立或不成立,因此需要根据假设进行相关数据的采集。针对上述假设,该糖果生产商可以组织一场试吃测试,让 50~100 个小朋友各自独立试吃两种不同配方的糖果,并记录他们更喜欢的糖果种类。

3. 分析并得出结论

选择合适的检验统计量对样本进行分析,并设定检验结果的显著性水平,根据显著性水平判断是接受还是拒绝原假设。如本例中数据是二项分布的,可以采用卡方检验来进行分析,并将显著性水平设定为 0.05,计算所采集样本数据检验统计量的概率 p 值,如果 p 值小于 0.05,则拒绝原假设,认为有足够证据支持新配方更受小朋友的欢迎,否则接受原假设,认为新配方和传统配方的受欢迎程度没有差别。

4. 作出决策

根据之前得出的结论作出最优决策。如果原假设被接受,则该糖果生产商要考虑是否在改进后再根据新的测试结果决定推广与否,如果原假设被拒绝,则该生产商可以考虑加大新配方糖果的推广力度,以提升销售量获得更多利润。

从上述步骤不难看出,提出假设是使用假设检验分析法时非常关键的一环。一般情况下,人们习惯于根据自身以往的经验提出假设,但这是存在风险的,很有可能会遗漏一些重要的假设。因此,如何客观提出假设就显得尤为重要。以“某品牌手机销售额下降”为例,如果要进行假设检验分析,可以从多个不同的维度考虑提出假设,如从用户维度提出类似于“用户更换手机的周期延长了”的假设,从产品维度提出类似于“手机的设计不符合用户需求”的假设,从竞争对手维度提出类似于“具有相同功能的竞品价格更低”的假设,通过假

设检验的分析结果可以快速定位造成问题的关键原因,根据假设所对应的具体部门,将后续的问题解决落实到各个部门。

1.5.6 归因分析法

归因分析法是一种综合性强、应用范围广的数据分析方法,它通过定量和定性的手段,对特定的结果进行解释和归因,是一种评估和分析多个因素对某一结果的贡献度的方法。通过归因分析找到事情发展的原因,识别所有对最终转化有贡献的因素,并确定每个因素的贡献度。归因分析法大致可以分成以下两种类型。

1. 单触点归因

单触点归因(single touchpoint attribution, STA)的核心思想是将转化的全部贡献都归功到某一因素上,常见的模型有如下两种。

(1) 首次触点归因模型:这种模型将对最终转化的贡献全部归于第一次的接触。例如某公司在进行品牌宣传时,更关心用户是在什么时间、什么地点第一次接触到该品牌,这有助于公司了解哪些渠道对于吸引用户更有效,为后续的市场拓展提供参考。

(2) 末次触点归因模型:这种模型将对最终转化的贡献全部归于最后一次的接触。例如某手机品牌商在进行市场调研时,更关心促使用户最后决定购买该品牌手机的宣传方式或渠道,这有助于该品牌商更有针对性地制定销售策略。

2. 多触点归因

和单触点归因不同,多触点归因(multi-touchpoint attribution, MTA)认为转化的贡献是多个不同因素共同作用的结果,而不会把这个贡献全部归因于某一个因素,常见的模型有如下三种。

(1) 线性归因模型:这种模型认为转化过程中的每一个接触都很重要,它们对最后转化的贡献是一样的。例如,一个用户在决定购买某品牌手机前,分别从电视广告、社交媒体推广、车身广告三个不同渠道接触到该品牌手机的信息,线性归因模型会认为这三个渠道对这次销售的贡献是相同的,在分析时应当赋予相同的权重。

(2) 时间衰减归因模型:这种模型认为接触点越接近最终转化的时间节点,其贡献越大,因此会对最近的接触点赋予更大的权重。例如,一个用户从考虑买一款手机到最后实际购买,经过了一个月的时间,月初用户浏览了品牌商的在线商城,月中在社交媒体的推广上又看到了该品牌手机的广告,月末通过某在线商城搜索并完成了购买,时间衰减归因模型会认为某在线商城搜索的贡献最大,在分析时会赋予更大的权重,具体的权重可以通过线性衰减、指数衰减等方法进行计算。

(3) 位置归因模型:这种模型是根据接触点在转化过程中所处的位置来分配权重,一般情况下,首末位置会被赋予更大的权重。以上例来说,用位置归因模型分析时,品牌商的在线商城和某在线商城搜索会被认为作出了同样多的贡献,社交媒体的推广的贡献则偏少,在分析时,前两者可能会分别被赋予40%的权重,而后者则被赋予20%的权重。

1.5.7 预测分析法

对数据进行分析的目的不仅仅是了解过去,更重要的是预测未来的行为或趋势,以应对潜在危机、优化未来的工作等。预测分析法也是一种非常重要的数据分析方法,可以在结构化数据和非结构化数据(unstructured data)中使用,它是一种基于统计或数据挖掘技术,通过对过去和现在数据的分析来预测未来的过程。常用的预测分析法有如下三种。

1. 时间序列分析

该方法主要用于分析按时间顺序排列的数据,其核心在于理解和预测数据随时间变化的趋势与模式。该方法的应用非常广泛,如企业可以用来进行销售和 demand 预测,金融领域可以用来预测股票价格、汇率和利率的变化,气象学领域可以用来预测温度、降雨量和风速等。相关的方法有很多,如移动平均法、指数平滑法、ARIMA 模型等。

2. 回归分析

该方法主要通过建立自变量和因变量之间关系的数学模型来实现预测。无论是在学术研究还是在商业实践中,回归分析都是一种不可或缺的分析方法。例如,在市场营销领域,可以通过分析广告投入和销售额之间的关系,更好地规划广告预算和预测销售收入;在医学领域,可以通过分析患者的基因、生活习惯、疾病史等因素来预测某种疾病发生的概率,为医生制订个性化治疗方案、提升治疗效果提供依据。相关的方法也有很多,如线性回归(linear regression)、逻辑回归(logistic regression)、非线性回归等。

3. 机器学习方法

在大数据时代背景下,数据规模正以指数级速度急剧膨胀,传统数据分析方法已难以应对如此庞大的数据挑战,机器学习方法的出现为处理和分析海量数据提供了全新的解决思路,它不仅能够处理和分析大规模的数据集,还能从中提取有价值的信息,为决策提供支持。例如,在交通领域可以用于预测交通流量和拥堵情况,以优化交通管理和路线规划;在旅游领域可以用于预测热门景点重大节假日的客流量,以制订合理的资源分配方案和有效的安全管理措施。相关的方法有很多,如决策树(decision tree)、随机森林(RF)、支持向量机(SVM)等。

1.6 数据分析在现实领域的应用

数据分析作为处理和解读数据的关键技术,已经渗透到我们日常生活的各个领域。从金融市场走势的精准预测,到电商平台为消费者提供的个性化推荐,再到医疗领域的精准治疗,数据分析的应用领域正在不断拓宽,展现出强大的应用潜力和价值。本节主要介绍一些数据分析应用较为广泛且深入的领域。

1.6.1 商业领域

商业领域是数据分析应用较早且涉及业务较为宽泛的领域。数据分析能力目前已经

成为企业不可或缺的重要能力之一,是企业在激烈的市场竞争中取得优势的重要助力。

(1) 销售预测。通过对历史销售数据、市场趋势、季节因素等的分析,预测未来一段时间内的销售趋势和潜在需求,提前制定销售策略。

(2) 渠道优化。通过对线上渠道和线下渠道、直销渠道和分销渠道的销售数据分析,发现最佳销售渠道组合,优化销售网络布局。

(3) 绩效提升。通过对销售人员的业绩数据分析,评估销售人员的绩效表现,制订针对性的培训和激励措施,推动绩效提升。

(4) 客户细分。通过对客户购买记录、浏览记录等数据的分析,将客户划分成不同的细分市场,制定更精准的市场营销策略,提供更具个性化的产品和服务,提升客户的满意度和忠诚度。

(5) 产品线调整。通过各类产品或系列产品的销售状况及区域市场接受度分析,调整产品线策略,以更好地满足市场需求。

(6) 竞争策略制定。通过对竞争对手产品、价格、市场占比等数据的分析,了解竞争对手的优劣势,准确定位自身的市场地位及竞争优势,有针对性地制定竞争策略,提高市场占有率。

1.6.2 金融领域

金融市场的日益复杂和竞争加剧,使金融机构面临越来越多的挑战和机遇,数据分析作为一种强大的工具,为金融机构准确把握市场动态、洞察客户需求、优化业务流程,在激烈的市场竞争中脱颖而出提供了有力的支持。

(1) 风险管理。通过客户的借记卡和信用卡交易记录、手机定位数据、社交媒体活动等分析,进行更全面、准确的信用评估,提高风险管理能力,有效识别和预防欺诈行为。

(2) 投资策略。通过对历史交易数据的分析,识别市场中的交易信号和模式,制定合理的投资决策,帮助投资者抓住投资机会,提高投资的准确性。

(3) 个性化营销。通过客户的购买历史、偏好和反馈等的分析,了解客户需求和行为模式,制定精准营销策略,提供个性化的产品和服务推荐,满足客户的特定需求。

(4) 优化运营效率。通过业务流程数据的分析,找到流程中的瓶颈问题,提出改进措施,降低运行成本,提高运营效率。

(5) 产品开发。通过客户交易数据、市场趋势、竞争对手产品特性等的分析,识别客户的痛点和期望,有针对性地进行产品设计,使产品更贴合市场需求。

(6) 信贷决策。通过对潜在借款人的信用历史、财务状况、还款记录等数据的分析,建立风险评分模型,全面评估借款人的信用状况和还款能力,对借款人进行风险评级,在降低信贷风险的同时,提高信贷审批的效率和准确性。

1.6.3 制造业领域

制造业作为国民经济的支柱产业,面临市场竞争激励、成本压力增大、客户需求多样化等挑战,掌握和应用数据分析技术已经成为其提升竞争力、实现转型升级的必然选择。

(1) 供应链管理。通过供应商、订单、库存、运输等数据的分析,监控整个供应链的运行情况,及时根据问题调整策略,优化供应链的效率和稳定性。

(2) 质量控制。通过生产过程中数据的分析,实时监测产品质量,及时发现潜在质量问题并采取相应改进措施,降低次品率,降低售后维修和退货成本。

(3) 预防性维护。通过机器设备传感器数据的分析,预测设备故障的可能性,提前介入维护措施,避免因设备故障导致生产线停机,减少生产延误和维修成本,提高设备利用率和整体生产效率。

(4) 环境与安全监控。通过设备的噪声、振动水平等数据的分析,发现潜在安全隐患和风险点,及时处理异常情况,保障工作环境的安全和健康。

(5) 人员配置优化。通过员工工作状态数据的分析,全面了解员工的工作效率,根据生产需求灵活调整员工岗位,实现人力资源的最大化利用,减少员工冗余和冲突,提高团队协作效率。

(6) 能源管理。通过历史和实时能源相关数据的分析,掌握能源消耗趋势,预测未来能源消耗情况,识别能源浪费的环节,优化能源消耗策略。

1.6.4 医疗领域

医疗领域拥有包括患者病历、诊断信息、治疗过程、药物效果等海量的数据资源,它们是医疗质量和效率提升的关键,而数据分析正是解锁这些数据价值的关键工具。

(1) 诊断与治疗。通过患者医疗记录、检查结果、影像学报告等数据的分析,医生能够更准确地诊断疾病,制订更合理的治疗方案,同时根据患者治疗过程数据的分析,及时调整治疗方案,以达到更好的治疗效果。

(2) 资源配置。通过对设备、病床、医护人员等资源运转情况数据的分析,评估资源的供需平衡,优化医疗设施、床位和医护人员配置,确保资源空间分配合理,医护人员在高峰时段也能充分应对患者需求,减少患者等待时间,提高服务品质。

(3) 疾病防控。通过疾病流行病学、环境因素和社会经济因素等数据的分析,预测疾病暴发或流行的趋势,及时采取预防措施,减少疾病蔓延给社会带来的冲击。

(4) 药物研发。通过药物分子结构、药理学特性和临床实验结果等数据的分析,辅助研发人员预测药物的效果、副作用和安全性,选择最有希望的合成方案,加速研发进程。

(5) 慢性病管理。通过可穿戴设备的健康状况数据分析,及时根据患者健康状况调整生活习惯、饮食和运动方案,有效控制慢性病的发展,降低并发症的风险。

1.6.5 交通领域

交通的运行效率、安全性和服务质量对于推动城市的可持续发展至关重要,数据分析作为一种能够深度挖掘交通数据、揭示其内在规律的技术手段,正逐渐成为推动交通领域进步的关键力量。

(1) 道路规划。通过道路交通流量、车辆行驶轨迹等数据的分析,预测未来的交通状况,制订科学的交通规划方案,优化道路网络布局,提高交通系统的整体效率。

(2) 拥堵治理。通过交通流量、车辆行驶速度等实时数据的分析,及时发现交通拥堵情况,智能调整交通信号灯时序,优化道路通行能力,缓解拥堵现象。

(3) 事故防范。通过交通事故历史数据的分析,找出事故发生的规律和原因,针对事故多发路段,采取加强交通监管、增设安全设施等措施,降低事故发生的概率。

(4) 公共交通管理。通过乘客出行需求、上下车地点和时间分布等数据的分析,了解乘客出行偏好和热点区域,优化公交车运营路线和发车间隔,满足乘客出行需求。

(5) 车辆监控。通过车辆行驶数据的实时监控和分析,及时发现超速、急刹车等危险行为,通过预警系统提醒驾驶员注意安全,提高道路安全水平。

1.6.6 政务领域

政务领域决策的科学性、管理的精细化和服务的优质化直接关系到国家发展和民生福祉,数据分析的应用也日益受到重视。

(1) 政策制定与评估。对社会经济发展现状、民生需求、产业结构等数据的分析,为政策制定提供科学依据,如通过税收数据评估税收政策的效果,根据分析结果调整税收率,这种基于数据驱动的决策方式,使得政策的制定更为精准和有效。

(2) 城市管理。通过历史和实时数据的分析,预测未来城市的人口分布、交通流量、能源消耗等情况,制订科学、合理的城市规划方案,优化城市空间布局,提高城市资源利用率。

(3) 应急管理。通过各类安全事件数据的分析,预测潜在的安全风险,制订预防措施,在应急事件发生时,帮助政府迅速了解事件情况,制定应急响应策略,降低损失。

(4) 财政监管。通过财政收支、税收、社保等数据的分析,了解财政资金流动情况,发现资源配置的不合理之处,优化财政支出结构,提高财政资金使用效率。

(5) 灾害预警。通过历史灾害数据的分析,识别灾害发生的模式、趋势和规律,预测未来可能发生的灾害类型和程度,制定预警机制,为灾害应对争取宝贵时间。



思考题

1. 简述数据分析的定义,并阐释其在现代社会中的重要性。
2. 数据分析的基本流程包括哪些步骤?
3. 常用的数据分析方法有哪些?



即测即练

