数据分析入门解析（一）

目录

备注：部分内容来源于网络，如有侵权请私信管理 Vivian：wmyd80，谢谢

图片来源于网络

一、概述

1、 什么是数据分析?

数据分析是一项从自然环境、社会环境，网络环境中提取数据，实施分析得出结论并实现业务价值。具体的说，就是用适当的统计分析方法将收集来的大量数据进行分析，将他们加以汇总和理解并消化，以求最大化地开发数据的功能，发挥数据的作用。为了提取有用信息和形成结论而对数据加以详细研究和概括总结的过程。

2、 数据分析常见类别

1）描述性数据分析

初级数据分析，常见的方法有对比分析法、平均分析法和交叉分析法。

2）探索性数据分析

侧重于在数据分析中发现新的特征，常见的方法有相关分析、因子分析和回归分析等。

3）验证性数据分析

侧重于检验已有假设的真伪证明，常见的方法有相关分析、因子分子和回归分析等。

3、 数据分析的作用

1）分析现状

表现形式为日常周报来完成，如日报、周报和月报等形式。

2）分析原因

知其然更要知其所以然，分析为什么事件会发生。表现形式专题分析，根据企业运营情况选择针对某一现状进行原因分析。

3）分析预测

将来时，预测未来会发生什么。表现形式专题分析，通常在制定企业制度、年度等计划时进行，其开展的频度没有现状分析和原因分析高。

4、 常见数据分析步骤

1）业务理解：确定目标对象，确认业务需求。

2）数据理解：根据业务对象，选择适合的数据维度，并确定其数据维度指标定义。

3）数据准备：通常所消耗物力、精力较大，是分析模型的基础。主要有数据获取，数据清洗，统计分析等。

4）建立模型：将数据以统计分析或机器学习算法对数据建模，以便描述数据或对未来进行预测。

5）结论价值：将数据分析转化成结论报告，并为业务提供指导支持。

二、数据获取的途径

我们根据把数据获取途径分别把优势和局限性分别对比：

1、公共数据库

优势：数据量大，使用免费，数据性权威

局限：数据粒度粗，数据更新慢，数据覆盖度低

2、私有数据库

优势：数据更新及时，数据粒度细

局限：数据价格高，数据访问权限受限

3、网络爬虫

优势：数据免费，数据量来源广
局限：技术要求高，数据质量差，可靠性低

4、问卷调查
优势：数据针对性高，数据可靠性高
局限：数据量少，数据范围小

5、设备采集
优势：数据准确度高
局限：获取成本高，数据范围小

三、数据清洗原则

为什么要做数据清洗，刚刚我们提及的，由于我们获取的渠道不同，所以数据质量残次不齐，拿到数据的第一步就是进行数据清洗，否则后续数据分析结果，模型结论都是偏差的。
数据清洗的主要原则：完整性、唯一性、一致性、合法性。
1、　完整性
主要是对于数据空字段进行修复补全，常见的操作手段一个是补全，一个是删除。补全信息一般是通过前后数据或者通过其他信息补全。如果涉及数据空字段较多，删除时候需要考虑其数据体量。

2、　唯一性
常见的就是同一个数据维度有两条数据，需要进行去重合并处理。但是向客户性别的这样唯一性的数据特性，肯定是无法合并的。这个时候就需要我们去看哪个数据获取来源更权威。

3、权威性

就是上边说的最权威的那个渠道的数据。

4、一致性

渠道不同，所以数据指标维度不同。现在公司一般都是搭建自有的数据体系，数据维度指标进行定义，包括维度、频度、计算方式等。

5、合法性

这里说的合法性有两个含义，一个是数据获取的渠道是合法的，在法律规定范围内。另外一个是数据定义是合法的。比如说日期表述方式可以是 2020-9-10，也可以是 2020 年 9 月 10 日。虽然在表述上都没错，但是在后续分析使用时候容易报错。

常见的数据合规规则有：字段内容、字段格式、离群值人工特殊处理。离群值人工特殊处理是指那些在数据中那些过大过小的特殊数据，作为孤立数据样本，一定要处理掉，否则直接影响模型准确度。如申请人的月收入普遍都是在 3000-10000 之间，这时候忽然冒出来一个 100 万的。

四、数据可视化解析

1、什么是数据可视化

数据可视化，是关于数据视觉表现形式的科学技术研究。其中，这种数据的视觉表现形式被定义为，一种以某种概要形式抽提出来的信息，包括相应信息单位的各种属性和变量。（来源于网络）可视化可简明地定义为：通过可视表达增强人们完成某些任务的效率。

2、为什么要做数据可视化

可视化后的数据不会超过数据本身，起码从信息层面是不可能超过的，但是能将不可见的数据现象转化为可见的图形符号，能将错综复杂、看起来没法解释和关联的数据，建立起联系和关联，发现规律和特征，更好的挖掘其数据背后的价值。

准确而高效、精简而全面地传递信息和知识。

3、常见的数据可视化类型

1）交通数据

常见的旅游，航空，物流等。如应用在春运时期检测人口流动规律，从而指定相应措施，配合返城潮。

2）地理信息

常见的区域地区对比，如人口、经济、温度。常见的每天七点北京电视台的电气预报。

3）数量对比

常见的如客群分析，如行业、收入等，一般用于饼状图，柱状图表示。

4）时间序列

常见股票分析图，按照时间走势分析数据。

5）多维度展示

如芝麻评分的五个维度，常见于申请中客户画像分析。

4、数据可视化的工具

1）通过工具

EXCEL：零门槛，适合小白也适合大神级

R 和 Python：灵活性高，大量的内嵌图标和三方库，比较美观

2）专业工具

Tableau：数据图表制作能力强，操作简洁不用写代码，数据导入和加载都是引导式下一步的。内置可视化图表，不用考虑配色，表格直接处理格式即可。

DataV：收费，编程简单，支持多种数据介入当时，大量的图标选择，有动画效果。

5、怎么做数据可视化?

1）明确问题

首先要明确图表要说明什么问题，需要展现的业务问题是什么，背后的业务逻辑是什么，以及相应的数据分析结果。

2）确定维度

明确数据的关系和需要对比的数据维度，是时间趋势还是分布状态，是对比维度（时间，同比，环比）还是空间维度（一线城市、二三线城市）。

3）选择图表

不同的图表都有其自身的优点和局限性，根据数据类型以及确定的数据维度选择适合的图表。

4）调整图表

验证图表结论是否和预期一致，调整坐标轴、颜色取值、图上标签等细

节。如有业务需要在恰当处备注文字说明。