

基于大数据分析的泰坦尼克号乘客生存预测软件设计

1 绪 论

1.1 大数据的现状与前景

大数据是指超出以往概念量级能形容的数据，以以前的技术无法进行完美或者快速的处理的数据集。掌握了海量的数据就代表掌握了大量的资本，这种资本不是传统意义上的金钱而是一种信息资本。要处理海量级信息数据对技术是一种考验，而且对拥有者的洞察力有较高的要求。下面将从三个方面分析大数据到底是什么：

第一个方面是理论，通过理论可以让我们产生认识然后再进行传播。为了更好地了解行业内有关大数据的描述和定性分析，我们要定义大数据的特征。这里不仅要简单的讨论大数据带来的好处，更要深入分析大数据为什么有这些价值，分析大数据未来的发展趋势，从人类的视角我们更在意大数据在个人隐私方面的特殊作用。

第二个方面是技术，技术可以更好地体现大数据的作用，技术进步同时大数据也会进步。例如现在更多地云计算分布式处理技术、存储技术和感知技术就是技术在大数据上的合理体现，实现了数据的采集、处理、存储、结果；

第三个方面是实践，在生活的方方面面都需要用到数据，比如互联网、政府、企业和个人都可以使用就是大数据在生活 and 科研中的具体实现，他们的出现极大的改善了我们的生活质量，让数据更精准，人们更方便，相信在未来大数据能发挥更大的作用。

大数据之所以大，只要体现在数据量上。将人类有印刷术以来的数据进行汇总大概有200PB(1PB=210TB)，但是如果不仅仅局限于纸质上的资料那么这股数据将达到5EB(1EB=210PB)，这个数据量才称得上真正的大。现今社会硬盘的大小从最开始的KB级达到如今的TB级，而企业需要处理的数据更为庞大TB级已经不足以满足，已经达到EB级。而且数据类型逐渐变多，数据不再仅仅是数字、文本这种结构化数据，非结构化数据的存储以及使用越来越频繁，比如图片、定位、音视频等随着科技的发展越来越需要存储，而处理这种复杂的、层次多的数据也对数据处理技术提出了新的挑战。但是数据越来越大还会造成一种问题——数据臃肿，以一段音频举例：这段音频3分20秒，可是我们只需要对其中间1分10秒到2分43秒的部分进行处理。以上这种现象不仅存在在音频中，视频中也很常见，所以这就需要我们使用处理数据能力更强大的数据对这些臃肿的数据进行“减重”

处理或者又叫做提纯。如何能更快更好的实现高效率的数据处理是现今背景下急需解决的问题。随着大数据概念的普及，在生活里使用的越来越频繁，我们需要意识到处理数据的效率至关重要，“时间就是金钱”。

关于大数据的未来前景，相信在未来大数据在我国的发展不可估量，就像互联网当初在中国发展的一样，甚至可以带来新一轮的技术革命，例如数据挖掘、机器学习和人工智能技术等，除了带来技术上的革命，还有可能发展成一门专门的学科被越来越多的人研究。新学科的设立还会增加一系列新的相关工作，未来还会通过大数据实现搭建跨域数据共享平台，更多地企业可以通过这个平台实现数据共享并成为未来行业的核心大链接数据云,云处理大型数据提供了灵活性,扩大基础设施,是大数据的一个平台。不同领域尝试将大数据与各种技术相结合，通过两种技术的优势互补使大数据产生一种新的形态，更快更好地推动未来的科技革命，使大数据能发挥更多的作用，方便人们的生活。

大数据时代已经到来，根据 IDC 和 Mckinsey 的大数据研究报告。大数据挖掘商业价值的方法主要分为以下几种：

- (1) 对顾客进行分类，不同顾客消费的侧重点不同，通过大数据对顾客进行划分然后“对症下药”采取相应的行动。这样既会为顾客提供便利，也减少了企业资源的浪费；
- (2) 大数据的出现实现了用技术发掘新需求；
- (3) 产业链之间的联系会更加的亲密，因投资风险的损失会减小，而回报率会变高；
- (4) 产品和服务都会得到更好地改进。

从理论上分析大数据可以应用在所有的产业中，受益面也会很广，但是发展较快的应该是第三产业。所谓的第三产业多指服务业，服务业通过大数据实现对顾客的分析，加强顾客对企业的粘合性、规避风险、实现低投入高回报。所以从以上产业中我们可以看出在大数据时代，数据代表价值，从数据中可以挖掘到价值。生活中有许多这样的例子，比如银行通过大数据可以实现合理地规避风险；企业通过大数据分析可以支出的合理分配，用更少的支出得到更高的回报；政府通过大数据分析可以实现合理地税收；大数据用在交通分析可以实现快速出行，选择最优路线等等。

1.2 本论文的结构

本次毕业设计需要使用 Anaconda3 软件。首先使用 Python 语言实现对训练集和数据集进行总览，再对缺失值进行处理，使用 matplotlib.pyplot 库对数据关系进行分析，筛选出重要的特征值。对缺失的特征值进行抛弃或预测，不断的调试所采用的方法并对

结果进行预测与验证。

本论文由五个部分组成。第一部分是整篇论文的绪论，主要是对大数据技术的现状与前景进行简单的科普以及介绍；第二部分是开发环境以及开发工具的介绍，主要讲述了本次设计所用到的语言和软件；第三部分是库函数，首先介绍软件所使用的库函数，之后再介绍库函数中封装的部分方法的作用；第四部分是算法的实现，主要讲述了如何对数据进行处理并完成预测；第五部分是运行结果，展现设计完成的软件后的运行结果。

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：

<https://d.book118.com/927161054016006141>