# DATA71011理解数据及其环境

# 课程作业项目

本课程项目主要涉及零售行业中不同商店的销售预测问题的数据预处理。该任务涉及分析从欧洲一家大型连锁药店收集的历史销售数据（Logo 课程内容包括：在德国的一家连锁店开展业务（英文）。目的是向您展示现实的商业案例，并了解和洞察一些可以充分准备数据以优化其分析价值的方法。

**评估和提交**

* 您团队的分析计划的形成性演示幻灯片应在截止日期（ 2024 年 12 月10日23:59 ）前通过电子邮件发送至[pradyumn.shukla@manchester.ac.uk](mailto:pradyumn.shukla@manchester.ac.uk) 。 请注意，每个小组必须 制作一段五分钟的视频，详细介绍分析计划。该视频可以上传到您的 Teams 频道。这将是一个评估部分（有关更多详细信息，请参阅手册）。
* 个人报告提交截止日期： 2025年1月26日23:59 。

请参阅以下页面中的更多要求和提示。

**业务背景描述**

准确预测销售额是全球零售商面临的最困难的挑战之一，因为销售额受促销、竞争、节假日、季节性和地域性等多种因素的影响。在这个项目中，总体业务目标是预测德国各地 1,115 家药店 6 周的日销售额，因为可靠的销售预测使商店经理能够提高零售业务的整体生产力和盈利能力并提高客户满意度。



然而，这个销售预测问题的挑战在于考虑各种因素并处理历史记录中缺失的数据。因此，您需要对历史数据集进行全面预处理，以便进行可靠和准确的预测，以及主要的数据准备任务，如集成、可视化、清理和 转型，值得探索。下面简要介绍可用的数据集。

* **商店.csv**

该excel文件包含1115家药店的补充信息。

|  |  |
| --- | --- |
| 柱子 | 描述 |
| 店铺 | 匿名商店编号 |
| 商店类型 | 4 种不同的商店模式：a、b、c、d |
| 品种 | 分类级别：a = 基本，b = 额外，c = 扩展 |
| 比賽距离 | 到最近的竞争对手商店的距离（以米为单位） |
| 比赛开始日期 | 最接近的竞争对手开业的大致月份 |
| 比赛开始年份 | 最接近的竞争对手开业的大致年份 |
| 促销2 | 持续且连续的促销活动（例如基于优惠券的邮寄活动）：0 = 商店不参与，1 = 商店参与 |
| Promo2SinceWeek | 商店开始参与 Promo2 的日历周 |
| Promo2SinceYear | 商店开始参与 Promo2 的年份 |
| 促销间隔 | Promo2 重新开始的连续间隔，指定促销活动重新开始的月份。例如，“二月、五月、八月、十一月”表示每轮基于优惠券的邮寄活动从该商店的任何一年的二月、五月、八月、十一月开始，因为优惠券（主要用于某些产品的折扣）通常有效期为三个月，并且需要在优惠券过期之前向客户发送新一轮邮件 |

* **训练.csv**

该文件包含历史销售数据，涵盖 2013 年 1 月 1 日至 2015 年 7 月 31 日的销售情况。它包括以下字段：

|  |  |
| --- | --- |
| 柱子 | 描述 |
| 店铺 | 匿名商店编号 |
| 星期几 | 星期几：1 = 星期一，2 = 星期二，… |
| 日期 | 给定日期 |
| 销售量 | 某一天的营业额 |
| 顾客 | 某一天的顾客数量 |
| 打开 | 指示商店当天是否营业：0 = 关门，1 = 营业 |
| 宣传 | 表示商店当天是否正在进行特定商店的促销活动 |
| 国定假日 | 表示国定假日。通常，除少数例外，所有商店在国定假日都关门。请注意，所有学校在公共假日和周末都关门。a = 公共假日，b = 复活节假期，c = 圣诞节，0 = 无 |
| 学校假期 | 表示（商店，日期）是否受到公立学校关闭的影响 |

* **测试.csv**

该文件与 train.csv 相同，只是 2015 年 8 月 1 日至 2015 年 9 月 17 日期间销售额和客户信息未知。

**预测准确性评估**

在本项目中，以下均方根百分比误差（RMSPE）或其他适当的误差 如果您希望在预处理的数据集上建立预测模型，可以用来评估预测准确性。

其中*N*是用于准确度评估的数据记录总数，是*第 i*条记录的实际销售额，是*第 i*条记录的销售预测。请注意，应适当处理零实际销售额。

|  |  |
| --- | --- |
| 一般要求：- | 您需要与小组成员合作，了解业务问题，制定数据预处理计划以获得形成性反馈，然后完成一份报告（长度：1500 字）进行个人评估。  您的工作应涵盖（但不限于）以下内容。  - 审查可用数据并根据其变量、质量和与销售预测的相关性进行描述，  - 根据需要将数据集链接在一起，  - 对数据进行适当的预处理以进行进一步的分析，例如，您可能想要对任何分类数据进行编码，创建新变量，确定有多少个缺失值并进行适当的处理等。  - 确定影响销售的关键因素，例如，您可能想检查竞争和促销是否会对销售产生影响，以及公共/学校假期如何影响销售波动。  -使用您确定的变量构建预测模型（可以是线性回归模型、神经网络模型或其他模型）。请确保证明您选择的建模方法的合理性。  -解释分析的关键结果、假设和局限性。  - 包括原始数据、项目实施和预测模型性能的适当可视化。证明可视化方法的选择是合理的。 |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

**撰写报告的一些技巧**

* 想象一下，你写这份报告是为了让人阅读，而不仅仅是为了让你通过课程！
* 报告应包括引言和结论。这两部分都有分数。
* 好的报告是一种叙述，而不只是简单地报告你所做的事情。
* 您的目标是传达您的发现，而不是简单地进行分析。
* 上述步骤是分析和报告中应包含的组成部分；如何包含它们由您决定。仅使用上述任务描述作为标题的报告将失去分数。
* 优秀级报告往往超出了规范要求——添加了上面未指定的额外想法、联系/分析或数据呈现方式。我赞成这些（只要它们做得好！），但它们并不是必需的。
* 在布局和展示上花点功夫——这些都是很容易做到的。
* 探索性分析应在适当的情况下包含在主报告中，并添加到叙述中。假设测试输出可以包含在附录中，任何探索性分析都可以包含在附录中，这些分析可以添加到您要讲述的故事中，但会使正文变得杂乱。
* 在图表和表格数量过少和过多之间找到适当的平衡。每页一到两张（取决于大小）是一个很好的经验法则。
* 你应该在结论中报告你所用数据的局限性，或者同一主题的未来研究可能需要寻找什么。
* 您应该充分且恰当地标注/编号图表。一般经验法则是，图表本身应该可以理解，而无需参考正文。在正文中，应使用“图 n”或“表 n”来引用图表，其中 n 是表格或图表在论文中的编号。请注意，“表”和“图”这两个词的首字母要大写（因为“表 1”是代词）。
* 任何抄袭来源/参考资料或其他团体作品的行为都将受到惩罚，并可能导致零分（请参阅您的课程手册）。
* 在截止日期前向 Blackboard提交本课程的课程报告。

下表列出了分数的指示性细目

|  |  |
| --- | --- |
| 评估报告​ | ％ |
| 介绍 | 10 |
| 方法论（主要数据预处理任务） | 二十五 |
| 结果（描述、讨论、分析等） | 二十五 |
| 可视化（选择、评估、清晰度） | 20 |
| 结论、启示和建议 | 15 |
| 布局和呈现 | 5 |