课程大作业：Kaggle 沃尔玛销量预测

大作业概述

本大作业要求同学们完成 Kaggle 平台上的沃尔玛销量预测比赛：使用沃尔玛的分级销售数据，来预测未来 28 天的每日销售额。同学们需要完整的经历问题分析、数据分析和处理、模型设计等步骤来尽可能的提升预测的准确程度。

任务介绍

（1） 阅读所附的文献，进行文献综述和算法总结。

（2） 在 kaggle 平台上参加沃尔玛销量预测比赛 https://www.kaggle.com/competitions/m5-forecasting-accuracy。

（3） 可以参考 kaggle 平台上其他队伍的 code 和思路，例如：https://www.kaggle.com/code/robikscube/m5-forecasting-starter-data-exploration 。

具体要求

请仔细阅读下述大作业的具体要求，并遵照要求完成大作业。

（1）阅读所提供相关论文，进行文献综述

【基本要求】20 分

阅读所提供有关表格数据（Tabular data）模型设计的参考文献并进行文献综述。

在进行文献综述时，**可以以时间为序**，根据相关领域的发展脉络，介绍各个时期的典型文献以及他们所提的主要方法的概述（基本原理、优缺点等）；**也可以以类别为纲**，结合相关领域的不同典型方法，分别介绍相关方法的基本原理、代表性文献、方法的优缺点等。

【加分项】≤10 分

梯度提升决策树（Gradient Boosting Decision Tree, GBDT）作为一种非深度学习方法，对表格数据的建模能力很强。在表格数据上，不仅是 GBDT，其他传统机器学习算法的效果有时也能够超过深度学习方法的效果，请分析原因。（可以从现象、原理、数据量、数据特点

等角度分析）

（2）算法和实验

2.1 问题和评价指标

【基本要求】10 分

对任务中所要解决的问题给出形式化的定义。比如通过形式化的数学方法给出相关任务的输入、输出的定义，相关任务的概率函数表达；假设使用神经网络解决该问题，神经网络建模函数的意义、神经网络的训练目标（损失函数）的定义等。

2.2 数据分析和处理（特征工程）

【基本要求】20 分

这一步的目的是清洗数据并将给出的原始信息处理成更好表达问题本质的特征，从而更容易让模型学习。

（1） 对数据集进行简要的描述。

（2） 进行数据清洗和观测，对数据缺失值和异常值进行处理。

（3） 区分连续和离散两种属性，分别进行特征的预处理（例如标准化、归一化等）。

（4） 观察特征的分布和特征对预测目标的重要性（相关性分析、卡方检验、互信息等方式）。参考重要性进行特征选择。

【加分项】≤ 5 分

使用组合变换、特征交叉等方法构造新的特征，并分析构造特征的原因、特征重要性和新特征带来的整体效果增益。

2.3 模型设计和超参数搜索

【基本要求】20 分

（1） 对使用的模型和算法进行详细的介绍和形式化描述，应当包括模型结构、损失函数等。

（2） 介绍影响算法效果的超参数有哪些，为什么这些超参数会对结果产生重要影响。

（3） 介绍使用的超参数搜索方法以及最后搜索到的最优超参数。

2.4 给出实验结果并进行分析

【基础要求】10 分

（1） 明确沃尔玛销量预测中的评价指标（Evaluation Metrics），在实验报告中给出该评价指标，并对其进行解释说明。

（2） 按照上述的特征工程和模型设计进行实验，给出实验结果并进行分析。

【加分项】 ≤ 5 分

对比观察多种模型，讨论效果差异产生的原因。

（3）实验报告与海报展示

3.1 撰写实验报告

【实验报告】10 分

（评价实验报告撰写是否规范、内容是否全面丰富、逻辑是否清晰、重点是否突出）

（1） 实验报告可以以中文撰写、也可以以英文撰写。要求重点突出、逻辑清晰。

（2） 实验报告的格式参考正式的 paper，建议包括：

报告题目：沃尔玛销量预测

个人信息：包括小组成员的姓名及学号；具体专业方向（不能只是电子信息）；电子邮箱

中文摘要及关键词

英文摘要及关键词

引言

1. 文献综述（这里 1 为建议编号，下同）（可细分为子章节，下同）

2. 问题分析和评价指标

3. 数据分析和处理

4. 模型设计

5. 超参数搜索

6. 实验结果及分析

7. 结论

8. 所完成的加分项（以表格方式给出所完成的加分项，并给出实验报告中的对应子章节索引）

9. 成员分工及贡献比（以表格方式给出，可以按照具体要求中的项目划分，也可更加细分）

10. Kaggle 比赛提交记录截图及每两次提交之间的改进

11. 心得体会

参考文献

附录：给出包括实验报告在内的大作业相关文件清单及相应说明 （即上传到网络学堂的文件内容；代码可放于一个目录，并对该目录作说明）

(3) 实验报告的表格、图片等要给出相应的表题、图题，并顺序编号，并在正文中相应地方给出引用；参考文献应在正文中给出相应的引用。

3.2 准备海报展示

【海报展示】10 分

（老师/助教/同学互评的加权成绩：包括海报的美观度、工作亮点总结、汇报展示的效果等）

(1) 请每个小组准备一张海报，应当包括报告题目、小组成员姓名、学号、具体专业方向（不能只是电子信息）、电子邮箱等；

(2) 海报内容：除了基本算法/模型的介绍之外，应突出自己工作的亮点部分：可以是模型的亮点、实验结果的亮点、除了基本要求之外完成的加分项的亮点、甚至是实验报告撰写的亮点、实验结果呈现形式的亮点、心得体会的亮点等等，总之能够凸显自己工作特色的所有东西都可以作为亮点给出来。

(3) 完成大作业后，会花一次课程的时间，让大家在课堂上展示和介绍自己的海报（需打印海报）。

(4) 海报电子版需使用 pptx 格式准备，设置为 A0 大小。

(5) 海报电子版需在规定时间（具体时间请等待通知）之前上传到网络学堂。

（4）在截止日期前上传实验结果

将实验报告、海报电子版 pptx 文件、所复现代码以及相关说明文件（如代码运行环境需求说明、代码运行方法说明等），打包成一个 zip 文件上传到网络学堂。

请在截止日期前上传实验结果。否则将按以下公式扣分：

$$S' = S \times \min\left(0.85, 0.95^{D}\right)$$

其中，S'是迟交作业的评分，S是作业的原始得分，D是向上取整的迟交天数（超过 deadline 后即记为迟交一天）。例如：作业的 deadline 是 10 月 11 日，10 月 12 日补交的作业评分为原始作业得分的 85%，10 月 18 日补交的作业评分将被折合为原始作业得分的 69.8%。

有关资源

相关参考文献的 3 篇论文，提供清华云盘下载地址：（访问密码：bigdatathu）

https://cloud.tsinghua.edu.cn/d/31df6b8400a14c4a919c/

本次作业的参考文献在“Kaggle”子目录下。

其他大家关心的问题

Q1：训练所需的计算资源和时间

A1：推荐大家使用 kaggle 提供的环境和计算资源，加速器可以选择 GPU（Tesla-P100 16G）免费时长 40h/week。训练时间会与大家的模型复杂程度有关。大家也可以选择使用自己的服务器完成任务。

Q2：可以参考其他人的教程和 code 吗？

A2：当然可以，kaggle 上有丰富的思路的 code 可以参考，对新手的帮助很大。但是大家要避免完全 copy，大家所提交的代码或 notebook，助教会检查是否存在抄袭现象。

Q3：代码会查重么？实验报告会查重么？

A3：代码不会查重，但是实验报告会查重。大家可以参考和下载已有的开源代码来完成大作业，但是一定要在实验报告中、说明文档中注明代码的来源。另外，必须要好好学习所下载的代码，要按照上述要求把整个实验过程和原理在实验报告中写清楚，不能只是跑了一个代码得到一个结果而已。

Q4：听说文献综述会影响实验报告的查重率？

A4：文献综述不是把别的论文中的内容简单拷贝粘贴过来，而是需要用自己的语言来对所阅读的论文进行总结，从论文所解决的问题、所提的方法、方法的优缺点等角度进行总结，具体要求见前述说明。通过自己的语言来进行文献综述，对实验报告查重率的影响应该可以控制得很小。

Q5：我遇到了问题怎么办？

A5：请放松随意地在课程群里提问，会有助教进行回答。

Q6：我能做完全部的加分项吗？

A6：非常鼓励有兴趣的同学自行尝试加分项的内容，但加分项最多只能累计 20 分。