1. **The UCI census-income dataset**

网址：<https://archive.ics.uci.edu/ml/datasets/Census-Income+%28KDD%29>

census-income.data：199523 条训练数据

census-income.test：99762条测试数据，

包含40个特征，本数据集主要用于分类任务

MMOE多目标论文中针对本数据集的两组实验预测的目标：

（1）Task 1: Predict whether the income exceeds $50K;

Task 2: Predict whether this person’s marital status is never married.

(2) Task 1: Predict whether the education level is at least college;

Task 2: Predict whether this person’s marital status is never married.

样本样例：

38, Private, 6, 36, 1st 2nd 3rd or 4th grade, 0, Not in universe, Married-civilian spouse present, Manufacturing-durable goods, Machine operators assmblrs & inspctrs, White, Mexican (Mexicano), Female, Not in universe, Not in universe, Full-time schedules, 0, 0, 0, Joint one under 65 & one 65+, Not in universe, Not in universe, Spouse of householder, Spouse of householder, 1032.38, ?, ?, ?, Not in universe under 1 year old, ?, 4, Not in universe, Mexico, Mexico, Mexico, Foreign born- Not a citizen of U S , 0, Not in universe, 2, 12, 95, - 50000.

目前的目前特征的字段：

| 91 distinct values for attribute #0 (age) continuous

| 9 distinct values for attribute #1 (class of worker) nominal

| 52 distinct values for attribute #2 (detailed industry recode) nominal

| 47 distinct values for attribute #3 (detailed occupation recode) nominal

| 17 distinct values for attribute #4 (education) nominal

| 1240 distinct values for attribute #5 (wage per hour) continuous

| 3 distinct values for attribute #6 (enroll in edu inst last wk) nominal

| 7 distinct values for attribute #7 (marital stat) nominal

| 24 distinct values for attribute #8 (major industry code) nominal

| 15 distinct values for attribute #9 (major occupation code) nominal

| 5 distinct values for attribute #10 (race) nominal

| 10 distinct values for attribute #11 (hispanic origin) nominal

| 2 distinct values for attribute #12 (sex) nominal

| 3 distinct values for attribute #13 (member of a labor union) nominal

| 6 distinct values for attribute #14 (reason for unemployment) nominal

| 8 distinct values for attribute #15 (full or part time employment stat) nominal

| 132 distinct values for attribute #16 (capital gains) continuous

| 113 distinct values for attribute #17 (capital losses) continuous

| 1478 distinct values for attribute #18 (dividends from stocks) continuous

| 6 distinct values for attribute #19 (tax filer stat) nominal

| 6 distinct values for attribute #20 (region of previous residence) nominal

| 51 distinct values for attribute #21 (state of previous residence) nominal

| 38 distinct values for attribute #22 (detailed household and family stat) nominal

| 8 distinct values for attribute #23 (detailed household summary in household) nominal

| 10 distinct values for attribute #24 (migration code-change in msa) nominal

| 9 distinct values for attribute #25 (migration code-change in reg) nominal

| 10 distinct values for attribute #26 (migration code-move within reg) nominal

| 3 distinct values for attribute #27 (live in this house 1 year ago) nominal

| 4 distinct values for attribute #28 (migration prev res in sunbelt) nominal

| 7 distinct values for attribute #29 (num persons worked for employer) continuous

| 5 distinct values for attribute #30 (family members under 18) nominal

| 43 distinct values for attribute #31 (country of birth father) nominal

| 43 distinct values for attribute #32 (country of birth mother) nominal

| 43 distinct values for attribute #33 (country of birth self) nominal

| 5 distinct values for attribute #34 (citizenship) nominal

| 3 distinct values for attribute #35 (own business or self employed) nominal

| 3 distinct values for attribute #36 (fill inc questionnaire for veteran's admin) nominal

| 3 distinct values for attribute #37 (veterans benefits) nominal

| 53 distinct values for attribute #38 (weeks worked in year) continuous

| 2 distinct values for attribute #39 (year) nominal

1. **MovieLens**

MovieLens1M是电影推荐数据集，包含 3,900 部电影和 6,040 个用户记录。

对于每个用户-电影对，构建了两个任务。

任务 1 是一个二分类问题，用于预测用户是否会观看特定电影。

任务 2 是一个回归问题，用于预测用户对给定电影的评分（{1, 2, . . , 5}）。

100 万训练样本。

ratings.dat是评级数据，格式如下：

UserID::MovieID::Rating::Timestamp

- UserIDs 范围在 1 到 6040 之间

- MovieIDs 范围在 1 到 3952 之间

- 评级采用 5 星制（仅限全星评级）

- 时间戳以 time(2) 返回的纪元以来的秒数表示

- 每个用户至少有 20 个评分

users.dat是用户信息，格式如下：

用户ID::性别::年龄::职业::邮政编码

- 性别用“M”表示男性，“F”表示女性

- 年龄选自以下范围：

\* 1：“未满 18 岁”

\* 18：“18-24”

\* 25：“25-34”

\* 35：“35-44”

\* 45：“45-49”

\* 50：“50-55”

\* 56：“56+”

- 职业可从以下选项中选择：

\* 0：“其他”或未指定

\* 1：“学术/教育家”

\* 2：“艺术家”

\* 3：“文员/管理员”

\* 4：“大学生/研究生”

\* 5：“客户服务”

\* 6：“医生/保健”

\* 7：“执行/管理”

\* 8：“农民”

\* 9：“家庭主妇”

\* 10：“K-12 学生”

\* 11：“律师”

\* 12：“程序员”

\* 13：“退休”

\* 14：“销售/营销”

\* 15：“科学家”

\* 16：“个体户”

\* 17：“技术员/工程师”

\* 18：“商人/工匠”

\* 19：“失业”

\* 20：“作家”

movies.dat是电影信息，格式如下：

电影 ID::标题::流派

- 标题与 IMDB 提供的标题相同（包括

发行年份）

- 流派是管道分隔的，选自以下流派：

\* 行动

\* 冒险

\* 动画

\* 孩子们的

\*喜剧

\* 犯罪

\* 纪录片

\* 戏剧

\* 奇幻

\* 黑色电影

\*恐怖

\* 音乐剧

\* 神秘

\*浪漫

\* 科幻

\*惊悚片

\* 战争

\* 西方

1. **Amazon product data**

网址：<https://nijianmo.github.io/amazon/index.html>

<https://cseweb.ucsd.edu/~jmcauley/datasets.html>

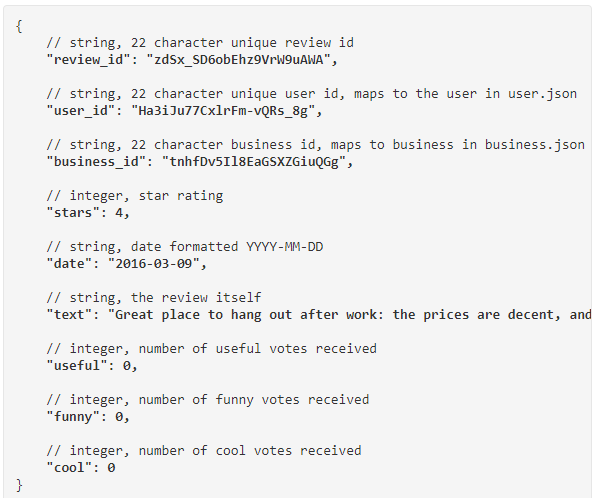
可以选取里面的两三个类别进行验证，用的比较多的是Electronics，Books， Movies and TV  
我们的预测目标可以分为是否点击与对商品的打分

特征分为两部分review(记录本次浏览的信息)和metadata（物品的特征）

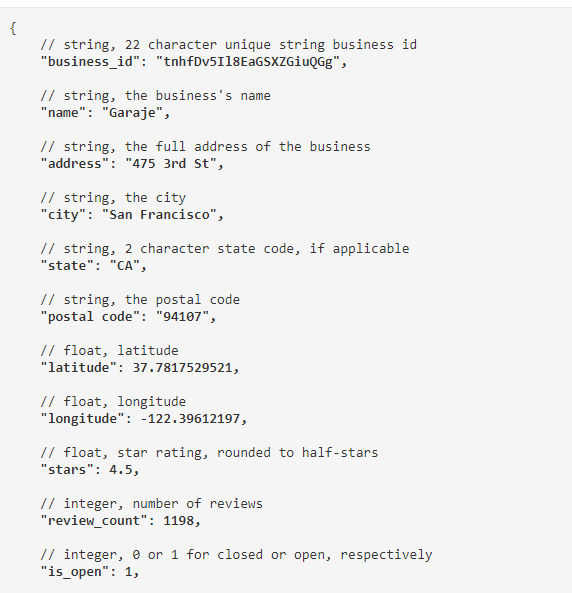




**4． Yelp Open Dataset  
网址：**[**https://www.yelp.com/dataset**](https://www.yelp.com/dataset)

可以预测的目标有：用户对商户的评分， 预测获赞等 **数据集包含：**（1） review.json 用户的浏览记录，对餐厅的评分和评价

（2）business.json



(3) user.json

